Biostatistics & Computer Applications

Estari Mamidala

FACULTY OF SCIENCE

M.Sc., ZOOLOGY

II Semester

Paper- IV

Biostatistics and Computer Applications

Unit-I Biostatistics

1. Introduction to Biostatistics – Definition, Terms, Applications and Role of biostatistics in modern research
2. Measures of central Tendency – Mean, Median, Mode, and Standard Deviation.
3. Student's t-Test, Chi-square test.
4. Correlation, Linear Regression and Logistic Regression.
5. Analysis of Variance – Types of ANOVA and classes of ANOVA models.

Unit-II Basics of computers

2.1 Basic Components of Computers – Hardware (CPU, input, output, Storage devices). Software (Operating systems).
2.2 Introduction to MS Excel – use of worksheet to enter data, edit data, copy data, move data in Graphical tools in EXCEL for presentation of data.
2.3 MS-WORD – editing, copying, moving, formatting, table insertion, drawing flow charts etc.,
2.4 Introduction to PPT, image, data handling and Graphical tools in PPT for presentation.
2.5 Use of Computers in Data Processing and Mapping.

Unit-III Internet Basics

3.1 Introduction to internet – Basics and Applications of Internet, Internet working, Internet access.
3.2 Using of Internet – understanding the basics, title , menu and tool bars, Address bar, Navigating web pages and Web sites and Printing.
3.3 Understanding the World Wide Web (WWW)
3.4 Searching Tools – World Search Engines, Search Directories and encyclopedias.
3.5 Online safety – Spywares and viruses.

Unit-IV Bioinformatics

4.1 Introduction, scope and applications of bioinformatics.
4.2 Biological Databases – Protein and DNA sequences data bases; importance.
4.3 Genomics – Definitions, Pharmacogenomics, taxicogenomics, human genomics, prokaryotic and eukaryotic genomes and genome relationships.
4.4 Proteomics – Definitions, Transcriptomics and Metabolomics, Proteomics techniques (2D PAGE)
4.5 Computational Biology – Multiple Sequence Analysis and Phylogenetic alignment.

Unit-I - BIOSTATISTICS

1.1 Introduction to Biostatistics – Definition, Terms, Applications and Role of biostatistics in modern research.
1.2 Measures of central Tendency – Mean, Median, Mode, and Standard Deviation. Student's t-Test, Chi-square test.
1.3 Correlation, Linear Regression and Logistic Regression.
1.4 Analysis of Variance – Types of ANOVA and classes of ANOVA models.

1.1 Introduction to Biostatistics – Definition, Terms, Applications and Role of biostatistics in modern research.

1.1.1 Definition:

- Biostatistics is the branch of statistics responsible for the proper interpretation of scientific data generated in the biology, public health and other health sciences (i.e., the biomedical sciences).
- In these sciences, subjects (patients, mice, cells, etc.) exhibit considerable variation in their response to stimuli. This variation may be due to different treatments or it may be due to chance, measurement error, or other characteristics of the individual subjects.
- Biostatisticians are specialists in the evaluation of data as scientific evidence. They understand the generic construct of data and they provide the mathematical framework that transcends the scientific context to generalize the findings. Their expertise includes the design and conduct of experiments, the mode and manner in which data are collected, the analysis of data, and the interpretation of results. Meaningful generalization of experimental results requires the application of an appropriate mathematical framework for the scientific context. The validity of research results depends on this application and the reproducibility of the experimental methods. Biostatisticians use mathematics to enhance science and bridge the gap between theory and practice.

1.1.2 Terms used in Biostatistics:

- ***Mean:*** equals the sum of observations divided by the number of observations.
- ***Median:*** equals the observation in the center when all observations are ordered from smallest to largest; when there is an even number of observations the median is defined as the average of the middle two values.
- ***Mode:*** equals the most frequently occurring value among all observations.
- ***Range:*** is the difference between the largest observation and the smallest.
- ***Standard Deviation:*** measures the spread of data around the mean. One standard deviation includes 68% of the values in a sample population and two standard deviations include 95% of the values.
- ***Standard Error of the Mean:*** describes the amount of variability in the measurement of the population mean from several different samples. This is in contrast to the standard deviation which measures the variability of individual observations in a sample.
- ***Percentile:*** equals the percentage of a distribution that is below a specific value. As an example, a child is in the 80th percentile for height if only 20% of children of the same age are taller than he is.
- ***Confidence Intervals:*** The results of any study sample are an estimate of the true value in the entire population. The true value may actually be greater or less than what is observed. A confidence interval gives a range of values within which there is a high probability (95% by convention) that the true population value can be

found. The confidence interval takes into consideration the number of observations and the standard deviation in the sample population. The confidence interval narrows as the number of observations increases or standard deviation decreases.

- ***Multiple linear regression:*** is used when the outcome data is a continuous variable such as weight. For example, one could estimate the effect of a diet on weight after adjusting for the effect of confounders such as smoking status. Another use of this method is to predict a linear variable based on known variables.

- ***Logistic regression:*** is used when the outcome data is binary such as cure or no cure. Logistic regression can be used to estimate the effect of an exposure on a binary outcome after adjusting for confounders. Logistic regression can also be used to find factors that discriminate two groups or to find prognostic indicators for a binary outcome. This method can also be applied to case-control studies.

- ***Clinical Trial:*** Experimental study in which the exposure status (e.g. assigned to active drug versus placebo) is determined by the investigator.

- ***Randomized Controlled Trial:*** A special type of clinical trial in which assignment to an exposure is determined purely by chance.

- ***Cohort Study:*** Observational study in which subjects with an exposure of interest (e.g. hypertension) and subjects without the exposure are identified and then followed forward in time to determine outcomes (e.g. stroke).

- ***Case-Control Study:*** Observational study that first identifies a group of subjects with a certain disease and a control group without the disease, and then looks to back in time (e.g. chart review) to find exposure to risk factors for the disease. This type of study is well suited for rare diseases.

- ***Cross-Sectional Study:*** Observational study that is done to ex-amine presence or absence of a disease or presence or absence of an exposure at a particular time. Since exposure and outcome are ascertained at the same time, it is often unclear if the exposure preceded the outcome.

- ***Case Report or Case Series:*** Descriptive study that reports on a single or a series of patients with a certain disease. This type of study usually generates a hypothesis but cannot test a hypothesis because it does not include an appropriate comparison group.

- ***ANOVA:*** Analysis of variance usually refers to an analysis of a continuous dependent variable where all the predictor variables are categorical. One-way ANOVA, where there is only one predictor variable (factor; grouping variable), is a generalization of the 2-samplet-test. ANOVA with 2 groups is identical to the t-test. Two-way ANOVA refers to two predictors, and if the two are allowed to interact in the model, two-way ANOVA involves cross-classification of observations simultaneously by both factors. It is not appropriate to refer to repeated measures within subjects as two-way ANOVA (e.g., treatment X time). An ANOVA table sometimes refers to statistics for more complex models, where explained variation from partial and total effects are displayed and continuous variables may be included.

1.1.3 Applications and Role of biostatistics in modern research:

Applications:

- Public health, including epidemiology, health services research, nutrition, environmental health and healthcare policy & management.
- Design and analysis of clinical trials in medicine
- Assessment of severity state of a patient with prognosis of outcome of a disease.
- Population genetics, and statistical genetics in order to link variation in genotype with a variation in phenotype. This has been used in agriculture to improve crops and farm animals (animal breeding). In biomedical research, this work can assist in finding candidates for gene alleles that can cause or influence predisposition to disease in human genetics.
- Analysis of genomics data, for example from microarray or proteomics experiments. Often concerning diseases or disease stages.
- Ecology, ecological forecasting.
- Biological sequence analysis.
- Systems biology for gene network inference or pathways analysis.

Role of biostatistics in modern research;

a) Biostatistics in Health

- For decades, biostatistics has played an integral role in modern medicine in everything from analyzing data to determining if a treatment will work to developing clinical trials. The University of North Carolina's Gillings School of Global Public Health defines biostatistics as "the science of obtaining, analyzing and interpreting data in order to understand and improve human health".
- Statisticians help medical researchers design studies, decide what data to collect, analyze data from medical experiments, help interpret the results of the analyses, and collaborate in writing articles to describe the results of medical research.
- To make it even plainer: biostatistics helps researchers make sense of all the data collected to decide whether a treatment is working or to find factors that contribute to diseases. Most biostatisticians have at least a master's degree and some have doctorates, often combined with master's degree in public health. Most majored in mathematics, statistics or computer science as undergrads. They work for pharmaceutical companies, universities and government agencies like the National Institutes of Health.

b) Biostatistics in epidemiology.

- When we hear statistics like one in eight women in the U.S. will develop invasive breast cancer over the course of her lifetime or that the risk factors for breast cancer are family history and age, we know that biostatics were instrumental in coming up with these conclusions.
- Epidemiology is the basic science of public health. It uses statistics and research methodologies to reach conclusions about diseases within certain population groups and finds the causes and risks of certain diseases.
- Biostatistics is used to determine how diseases develop, progress and spread. For example, biostatisticians use statistics to predict the behavior of an illness like the flu. It's used to help predict the mortality rate, the symptoms and even the time of year people might get it.
- Biostatics helps design the clinical trials to make sense out of the data, and help you draw the conclusion whether your home remedies will work or not.

c) Biostatistics in cancer research.

- Biostatistics is important in finding treatment for new drugs for diseases like cancer. "Cancer therapies tend to be very toxic. If you're a patient, and the usual therapy hasn't worked, you're desperate to find any therapy that offers hope to put you in remission.
- Biostatisticians look to design a study that tests as few patients as possible and gets them off the drugs quickly, so if they aren't working, they aren't subjected to the harmful side effects, he adds. All the while, the goal is to find which drugs work and ultimately reject the therapies that don't.
- Biostatisticians help design, manage and analyze cancer clinical trials. They also help identify the causes and characteristics of cancer. Oncologists rely on these numbers to recommend treatments for their cancer patients. Since cancer is not a "one-size fits all" disease, biostatisticians and oncologists work closely together to identify how factors such as drug interaction, diet and nutrition play a role in cancer. They also examine the traits of cancer and how it occurs in various ages, genders and racial groups to work on prevention and treatment.

2. Measures of central Tendency – Mean, Median, Mode, and Standard Deviation

2.1 Mean:

The mean is just the average of the numbers. It is easy to calculate: add up all the numbers, then divide by how many numbers there are.

Example 1: ***What is the Mean of these numbers? (6, 11, and 7)***

Solution: Add the numbers: 6 + 11 + 7 = 24
Divide by how many numbers (there are three numbers): 24/3 = 8
Therefore the **Mean is = 8**

Types of Mean:

In statistics, **mean** has two related meanings:

- the arithmetic mean (and is distinguished from the geometric mean or harmonic mean).
- the expected value of a random variable, which is also called the *population mean*.

As well as statistics, means are often used in geometry and analysis; a wide range of means have been developed for these purposes, which are not much used in statistics. These are listed below:

a) Arithmetic mean (AM)
b) Geometric mean (GM)
c) Harmonic mean (HM)

a) Arithmetic mean (AM)

The *arithmetic mean* is the "standard" average, often simply called the "mean".

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

The **mean** may often be confused with the median, mode or range. The mean is the arithmetic average of a set of values, or distribution; however, for skewed distributions, the mean is not necessarily the same as the middle value (median), or the most likely (mode).

For example, mean income is skewed upwards by a small number of people with very large incomes, so that the majority have an income lower than the mean. By contrast, the median income is the level at which half the population is below and half is above. The mode income is the most likely income, and favors the larger number of people with lower incomes. The median or mode are often more intuitive measures of such data.

Nevertheless, many skewed distributions are best described by their mean – such as the exponential and Poisson distributions.

For example, the arithmetic mean of six values: 34, 27, 45, 55, 22, 34 is

$$\frac{34+27+45+55+22+34}{6} = \frac{217}{6} \approx 36.167.$$

b) Geometric mean (GM)

The geometric mean is an average that is useful for sets of positive numbers that are interpreted according to their product and not their sum (as is the case with the arithmetic mean) e.g. rates of growth.

$$\bar{x} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}$$

For example, the geometric mean of six values: 34, 27, 45, 55, 22, 34 is:

$$(34 \cdot 27 \cdot 45 \cdot 55 \cdot 22 \cdot 34)^{1/6} = 1,699,493,400^{1/6} \approx 34.545.$$

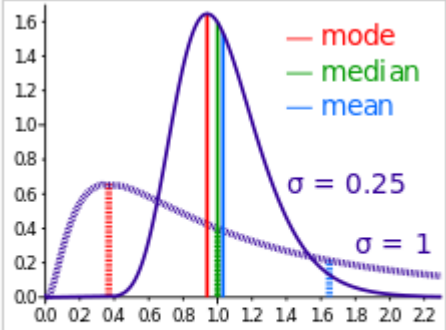
c) Harmonic mean (HM)

The harmonic mean is an average which is useful for sets of numbers which are defined in relation to some unit, for example speed (distance per unit of time).

$$\bar{x} = n \cdot \left(\sum_{i=1}^{n} \frac{1}{x_i}\right)^{-1}$$

For example, the harmonic mean of the six values: 34, 27, 45, 55, 22, and 34 is

$$\frac{6}{\frac{1}{34}+\frac{1}{27}+\frac{1}{45}+\frac{1}{55}+\frac{1}{22}+\frac{1}{34}} = \frac{60588}{1835} \approx 33.0179836.$$



Comparison of the arithmetic mean, median and mode of two skewed (log-normal) distributions.

2.2 Median:

The Median is the ***"middle number"*** (in a sorted list of numbers). To find the Median, place the numbers you are given in **value order** and find the **middle number**.

Example 1: ***Find the Median of (12, 3 and 5) ?***

Solution: Put them in order: 3, 5, 12
The middle number is 5, so the median is 5.

Example 2: ***Find the Median of the following numbers?***
3, 13, 7, 5, 21, 23, 39, 23, 40, 23, 14, 12, 56, 23, 29

Solution: If we put those numbers in order we have:
3, 5, 7, 12, 13, 14, 21, 23, 23, 23, 23, 29, 39, 40, 56
There are **fifteen** numbers. Our middle number will be the **eighth** number:
3, 5, 7, 12, 13, 14, 21, **23**, 23, 23, 23, 29, 39, 40, 56
The median value of this set of numbers is **23**.

2.3 Mode:

The mode is simply the number which appears **most often**. To find the mode, or modal value, first put the numbers **in order**, then count how many of each number.

Example: ***Find the Mode of the following numbers?***
3,7,5,13,20,23,39,23,40,23,14,12,56,23,29
Solution: In order these numbers are:
3,5,7,12,13,14,20,**23,23,23,23**,29,39,40,56
This makes it easy to see which numbers appear most often.
In this case the mode is **23.**

2.4 Standard Deviation:

The Standard Deviation is a measure of how spread out numbers are. Its symbol is σ (the greek letter sigma). A low standard deviation indicates that the data points tend to be very close to the mean, whereas high standard deviation indicates that the data points are spread out over a large range of values.

The formula is easy: it is the **square root** of the **Variance (**The Variance is defined as: The average of the **squared** differences from the Mean**).**

This is the formula for Standard Deviation:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

To calculate the standard deviation:

1. Work out the Mean (the simple average of the numbers)
2. Then for each number: subtract the Mean and square the result
3. Then work out the mean of **those** squared differences.
4. Take the square root of that and you are done!

Example: 1. ***Consider a population consisting of the following eight values:***

$$2,\ 4,\ 4,\ 4,\ 5,\ 5,\ 7,\ 9$$

Solution: These eight data points have the mean (average) of 5:

$$\frac{2+4+4+4+5+5+7+9}{8} = 5$$

To calculate the population standard deviation, first compute the difference of each data point from the mean, and square the result of each:

$$\begin{array}{ll} (2-5)^2 = (-3)^2 = 9 & (5-5)^2 = 0^2 = 0 \\ (4-5)^2 = (-1)^2 = 1 & (5-5)^2 = 0^2 = 0 \\ (4-5)^2 = (-1)^2 = 1 & (7-5)^2 = 2^2 = 4 \\ (4-5)^2 = (-1)^2 = 1 & (9-5)^2 = 4^2 = 16 \end{array}$$

Next compute the average of these values, and take the square root:

$$\sqrt{\frac{(9+1+1+1+0+0+4+16)}{8}} = 2$$

This quantity is the **population standard deviation**; it is equal to the square root of the variance.

*****

3. Student's t-Test, Chi-square test.

3.1 Student's T-test:

Student's t-test, in statistics, a method of testing hypotheses about the mean of a small sample drawn from a normally distributed population when the population standard deviation is unknown.

In 1908 William Sealy Gosset, an Englishman publishing under the pseudonym Student, developed the t-test and t distribution. The t distribution is a family of curves in which the number of degrees of freedom (the number of independent observations in the sample minus one) specifies a particular curve. As the sample size (and thus the degrees of freedom) increases, the t distribution approaches the bell shape of the standard normal distribution. In practice, for tests involving the mean of a sample of size greater than 30, the normal distribution is usually applied.

It is usual first to formulate a null hypothesis, which states that there is no effective difference between the observed sample mean and the hypothesized or stated population mean—i.e., that any measured difference is due only to chance. In an agricultural study, for example, the null hypothesis could be that an application of fertilizer has had no effect on crop yield, and an experiment would be performed to test whether it has increased the harvest. In general, a t-test may be either two-sided (also termed two-tailed), stating simply that the means are not equivalent, or one-sided, specifying whether the observed mean is larger or smaller than the hypothesized mean. The test statistic t is then calculated. If the observed t-statistic is more extreme than the critical value determined by

the appropriate reference distribution, the null hypothesis is rejected. The appropriate reference distribution for the t-statistic is the t distribution. The critical value depends on the significance level of the test (the probability of erroneously rejecting the null hypothesis).

For example, suppose a researcher wishes to test the hypothesis that a sample of size n = 25 with mean x = 79 and standard deviation s = 10 was drawn at random from a population with mean μ = 75 and unknown standard deviation. Using the formula for the t-statistic, the calculated t equals 2. For a two-sided test at a common level of significance α = 0.05, the critical values from the t distribution on 24 degrees of freedom are −2.064 and 2.064. The calculated t does not exceed these values, hence the null hypothesis cannot be rejected with 95 percent confidence. (The confidence level is 1 − α.)

A second application of the t distribution tests the hypothesis that two independent random samples have the same mean. The t distribution can also be used to construct confidence intervals for the true mean of a population (the first application) or for the difference between two sample means (the second application).

3.2 Chi-square test:

The Chi Square test is the most important and most used method in statistical tests. The purpose of Chi Square test is know as the difference between an observed frequency and expected frequency. This test, sometimes is also used to test the differences between the two or more observed data. Its value can be calculated by using the given observed frequency and expected frequency.

The Chi Square is denoted by X^2 and the formula is given as:

$$X^2 = \sum \frac{(O - E)^2}{E}$$

Here,
O = Observed frequency
E = Expected frequency
∑ = Summation
X^2 = Chi Square value

Chi Square Test Degrees of Freedom

The degree of freedom for the chi square difference test is equal to the difference between degree of freedom associated with the models. Each type of two way table has its own chi-square distribution, depending on the number of rows and columns, and each chi-square distribution is identified by its degree of freedom. A two way table with r rows and c column uses a chi-square distribution with (r - 1)*(c - 1) degree of freedom.

For one degree of freedom, the distribution looks like a hyperbola.
For than one degree of freedom, it loos like a mound that has a long right tail.

Chi Square Test of Independence

Chi square test is applied when we have two categorical variables from a single population. It is used to determine whether there is a significant association between the two variables. This test is applicable when the observations are independent (random). The Chi-square test for independence is also called a contingency table Chi-square test.

Example Question 1:

Find the chi square for the following given datas

Color	Blue	Black	Brown	Yellow
Observed frequency	5	15	10	20
Expected frequency	10	20	5	30

Solution:

Color	Observed frequency	Expected frequency	Observed frequency -Expected Frequency (O-E)	$(O-E)^2$	$\frac{(O-E)^2}{E}$
Blue	5	10	-5	25	2.5
Black	15	20	-5	25	1.25
Brown	10	5	5	25	5
Yellow	20	30	-10	100	0.83333

The formula for define Chi Square test is given by,

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

= (2.5 + 1.25 + 5 + 3.3333)

=9.58333

4. Correlation, Linear Regression and Logistic Regression:

4.1 Correlation:

Definition of correlation:

- The relationship between more than one variable is considered as correlation. Correlation is considered as a number which can be used to describe the relationship between two variables. Simple correlation is defined as a variation related amongst any two variables.
- The multiple correlation and partial correlation are categorized as related variation among three or more variables. Two variables are correlated only when they vary in such a way that the higher and lower values of one variable corresponds to the higher and lower values of the other variable. We might also get to know if they are correlated when the higher value of one variable corresponds with the lower value of the other.

Correlation Symbol = r

Correlation Formula:

The formula for correlation is as follows,

$$\text{Correlation (r)} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}}$$

Where, x and y are the variables.
b = the slope of the regression line is also called as the regression coefficient
a = intercept point of the regression line which is in the y-axis.
N = Number of values or elements
X = First Score
Y = Second Score
∑XY = Sum of the product of the first and Second Scores
∑X = Sum of First Scores
∑Y = Sum of Second Scores
∑X2 = Sum of square first scores.
∑Y2 = Sum of square second scores.

Types of Correlation:

There are different types of Correlation. They are listed as follows:

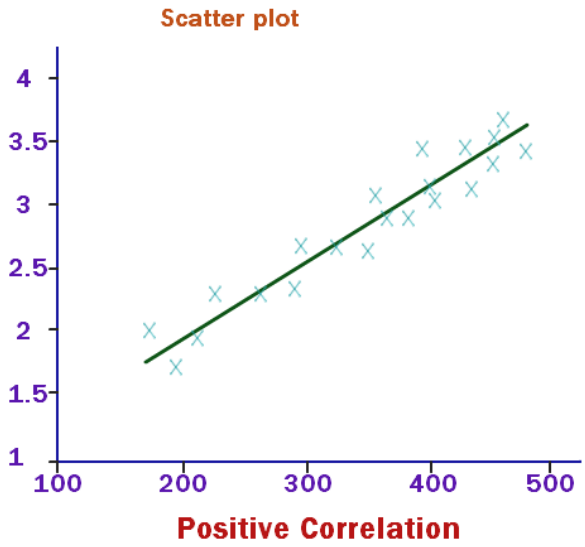
- **Positive Correlation:** A positive correlation is a correlation in the same direction.
- **Negative Correlation:** A negative correlation is a correlation in the opposite direction.
- **Partial Correlation:** The correlation is partial if we study the relationship between two variables keeping all other variables constant.
  - Example:
    The Relationship between yield and rainfall at a constant temperature is partial correlation.
- **Linear Correlation:** When the change in one variable results in the constant change in the other variable, we say the correlation is linear. When there is a linear correlation, the points plotted will be in a straight line.
  - Example:
    Consider the variables with the following values.

X:	10	20	30	40	50
Y:	20	40	60	80	100

    Here, there is a linear relationship between the variables. There is a ratio 1:2 at all points. Also, if we plot them they will be in a straight line.

- **Zero Order Correlation :** One of the most common and basic techniques for analyzing the relationships between variables is zero-order correlation. The value of a correlation coefficient can vary from -1 to +1. A -1 indicates a perfect negative correlation, while a +1 indicates a perfect positive correlation. A correlation of zero means there is no relationship between the two variables.

- **Scatter Plot Correlation :** A scatter plot is a type of mathematical diagram using cartesian coordinates to display values for two variables for a set of data. Scatter plots will often show at a glance whether a relationship exists between two sets of data. The data displayed on the graph resembles a line rising from left to right. Since the slope of the line is positive, there is a positive correlation between the two sets of data.



- **Spearman's Correlation:** Spearman's rank correlation coefficient allows us to identify easily the strength of correlation within a data set of two variables, and whether the correlation is positive or negative. The Spearman coefficient is denoted with the Greek letter rho (ρ).

  $\Rightarrow \rho = 1 - 6\sum d2n(n2-1)$

- **Non Linear Correlation:** When the amount of change in one variable is not in a constant ratio to the change in the other variable, we say that the correlation is non linear.

  Example:
  Consider the variables with the following values

X:	10	20	30	40	50
Y:	10	30	70	90	120

Here there is a non linear relationship between the variables. The ratio between them is not fixed for all points. Also if we plot them on the graph, the points will not be in a straight line. It will be a curve.

- **Simple Correlation :** If there are only two variable under study, the correlation is said to be simple.

Example:
The correlation between price and demand is simple.

Multiple Correlations:

When one variable is related to a number of other variables, the correlation is not simple. It is multiple if there is one variable on one side and a set of variables on the other side.

Example:

Relationship between yield with both rainfall and fertilizer together is multiple correlations

- **Weak Correlation:**

The range of the correlation coefficient between -1 to +1. If the linear correlation coefficient takes values close to 0, the correlation is weak.

4.2 LINEAR REGRESSION:

Introduction:

A regression is a statistical analysis assessing the association between two variables. It is used to find the relationship between two variables. Linear Regression Definition states that it can be measured by using lines of regression. Regression measures the amount of average relationship or mathematical relationship between two variables in terms of original units of data. Whereas, correlation measures the nature of relationship between two variables. i.e.., positive or negative or uncorrelated.

Definition of Linear Regression:

Regression is used for estimating the value of one variable if we know the value of other variable, one of the variable is independent variable and other variable is dependent variable.

Let (X_i , Y_i) ; i = 1, 2, 3, ...n the n pairs of observations are given now plot all these points in XY-plane which reserves a scatter diagram. In scatter diagram if the maximum number of points are going through a straight lines then we call it as linear regression if not that means they are passing through a curve then we call it as curve linear regression. Linear Regression can be measured by using lines of regression i.e.., Y-on-X & X-on-Y and also curve linear regression can be measured by using correlation ratio.

Linear Regression Equation

A linear regression line has an equation of the form Y = a + bX, where X is the explanatory variable and Y is the dependent variable, 'b' is the slope of the line, and 'a' is the intercept. The linear regression formula is derived as follows. Let (X_i , Y_i) ; i = 1, 2, 3,....., n be n-pairs of observations are given and there are representing a linear regression.

We know that, coefficient of correlation

$r = Cov(X,Y)\sigma X\sigma Y$

where Cov (X , Y) = $1n \sum XiYi - \bar{X}\bar{Y}$

and $X2=\frac{1}{n}\sum (X_i - \bar X)^2$

Now, we want to obtain regression equation of Y-on-X by taking the line and the corresponding normal equation are

Y = a + bX ---------- (1)

$\sum Yi = na + b\sum Xi$ ..... (2)

$\sum XiYi = a\sum Xi + b\sum X2i$ ..... (3)

Divide equation (2) and (3) by n

From (2), $1n\sum Yi=a+b(1n)\sum Xi$

$\bar{Y}=a+b\bar{X}$ ..... (4)

cov (X , Y) + X Y = $a\bar{X}+b(\sigma X2+\bar{X}2)$ ..... (5)

$\sigma X2=(1n)\sum X2i-\bar{X}2$

Multiplying equation (4) with X and sub from (5)

b = cov (X , Y) / σ X2

Substitute the value of b in (4)

Therefore, $Y-\bar{Y}=bYX(X-\bar{X})$

Similarly, we can prove that regression equation of X-on-Y is

$X-\bar{X}=bXY(Y-\bar{Y})$

4.3 LOGISTIC REGRESSION:

Definition of LR:

Linear regression is used to measure the degree of relationship between a dependent variable and a independent variable. If the response variable is binary, we need to consider using generalized linear models. One of the most common used generalized model is logistic regression. Logistic regression may be applied to a database where there is a nonlinear relationship between the response variable and one or more predictor variables. It is commonly used for predicting the probability of occurrence of an event, based on several predictor variables that may either be numerical or categorical.

Logistic Regression Assumptions

In logistic regression no assumptions are made about the distributions of the explanatory variables. However, the explanatory variables should not be highly correlated with one another because this cause problems with estimation.

Stepwise Logistic Regression

There are several procedure for variable selection implemented in statistics packages like backward elimination, forward selection, stepwise selection etc. Stepwise selection of variables is widely used in linear regression. Most of the software packages offer an option for stepwise logistic regression. Employing a stepwise selection procedure can provide a fast and effective means to screen a large number of variables, and to fit number of logistic regression equations. The result of stepwise logistic regression will depend substantially on the significance level for variable entering and staying in the model and selection criteria used.

5. Analysis of Variance – Types of ANOVA and classes of ANOVA models

5.1 Types of ANOVA

Definition & Introduction:

- Analysis of variance (ANOVA) is a collection of statistical models used to analyze the differences among group means and their associated procedures (such as "variation" among and between groups), developed by statistician and evolutionary biologist Ronald Fisher. In the ANOVA setting, the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether or not the means of several groups are equal, and therefore generalizes the t-test to more than two groups. ANOVAs are useful for comparing (testing) three or more means (groups or variables) for statistical significance. It is conceptually similar to multiple two-sample t-tests, but is less conservative (results in less type I error) and is therefore suited to a wide range of practical problems.

General Purpose of ANOVA

- These days, researchers are using ANOVA in many ways. The use of ANOVA depends on the research design. Commonly, researchers are using ANOVA in three ways: one-way ANOVA, two-way ANOVA, and N-way Multivariate ANOVA.

- ***One-Way:***
- When we compare more than two groups, based on one factor (independent variable), this is called one way ANOVA. For example, it is used if a manufacturing company wants to compare the productivity of three or more employees based on working hours. This is called one way ANOVA.

- ***Two-Way:***
- When a company wants to compare the employee productivity based on two factors (2 independent variables), then it said to be two way (Factorial) ANOVA. For example, based on the working hours and working conditions, if a company wants to compare employee productivity, it can do that through two way ANOVA. Two-way ANOVA's can be used to see the effect of one of the factors after controlling for the other, or it can be used to see the INTERACTION between the two factors. This is a great way to control for extraneous variables as you are able to add them to the design of the study.

Types:

- There are three types of ANOVA's that can candle an unbalanced design. These are the Classical Experimental design (Type 2 analysis), the Hierarchical Approach (Type 1 analysis), and the Full regression approach (Type 3 analysis). Which approach to use depends on whether the unbalanced data occurred on purpose.

    - If the data is unbalanced because this is a reflection of the population and it was intended, use the Full Regression approach (Type 3).
    - If the data was not intended to be unbalanced but you can argue some type of hierarchy between the factors, use the Hierarchical approach (Type 1).
    - If the data was not intended to be unbalanced and you cannot find any hierarchy, use the classical experimental approach (Type 2).

5.2 Classes of ANOVA models

There are three classes of models used in the analysis of variance, and these are outlined here.

a) Fixed-effects models

- The fixed-effects model (class I) of analysis of variance applies to situations in which the experimenter applies one or more treatments to the subjects of the experiment to see whether the response variable values change. This allows the experimenter to estimate the ranges of response variable values that the treatment would generate in the population as a whole.

b) Random-effects models

- Random effects model (class II) is used when the treatments are not fixed. This occurs when the various factor levels are sampled from a larger population. Because the levels themselves are random variables, some assumptions and the method of contrasting the treatments (a multi-variable generalization of simple differences) differ from the fixed-effects model.

c) Mixed-effects models

- A mixed-effects model (class III) contains experimental factors of both fixed and random-effects types, with appropriately different interpretations and analysis for the two types.

  Example: Teaching experiments could be performed by a university department to find a good introductory textbook, with each text considered a treatment. The fixed-effects model would compare a list of candidate texts. The random-effects model would determine whether important differences exist among a list of randomly selected texts. The mixed-effects model would compare the (fixed) incumbent texts to randomly selected alternatives.

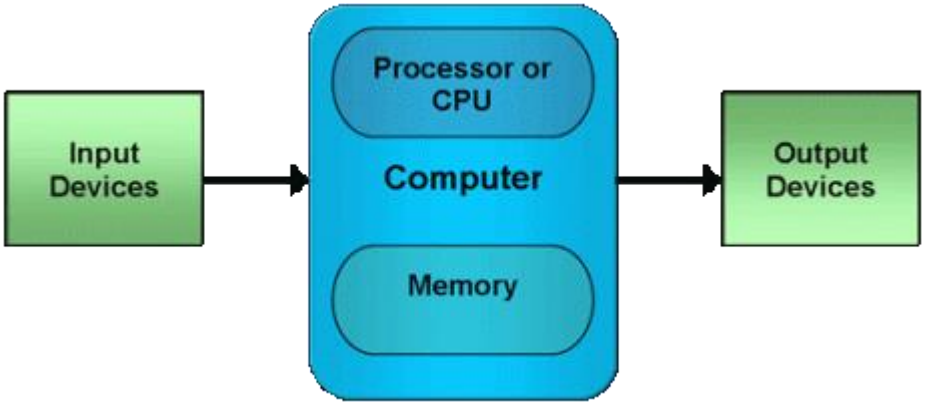
Defining fixed and random effects has proven elusive, with competing definitions arguably leading toward a linguistic quagmire

* * * * *

UNIT-II BASICS OF COMPUTERS

2.1 Basic Components of Computers – Hardware (CPU, input, output, Storage devices) and Software (Operating systems).

- The internal architectural design of computers differs from one system model to another. However, the basic organization remains the same for all computer systems. The following five units (also called "The functional units" or Hardware) correspond to the four basic operations performed by all computer systems.



2.1.1 Hardware (CPU, Input, Output & Storage devices):

- Hardware refers to the physical elements of a computer. This is also sometime called the machinery or the equipment of the computer. Examples of hardware in a computer are the keyboard, the monitor, the mouse and the processing unit. However, most of a computer's hardware cannot be seen; in other words, it is not an external element of the computer, but rather an internal one, surrounded by the computer's casing (tower). A computer's hardware is comprised of many different following parts:

- ***Central Processing Unit (CPU):***
- The main unit inside the computer is the **CPU**. This unit is responsible for all events inside the computer. It controls all internal and external devices, performs ***"Arithmetic and Logical operations"***. The operations a Microprocessor performs are called ***"instruction set"*** of this processor. The instruction set is “hard wired” in the CPU and determines the machine language for the CPU. The more complicated the instruction set is, the slower the CPU works. Processors differed from one another by the instruction set. If the same program can run on two different computer brands they are said to be compatible. Programs written for IBM compatible computers will not run on Apple computers because these two architectures are not compatible.
- The control Unit and the Arithmetic and Logic unit of a computer system are jointly known as the Central Processing Unit (CPU). The CPU is the brain of any computer system. In a human body, all major decisions are taken by the brain and the other parts of the body function as directed by the brain. Similarly, in a computer system, all major calculations and comparisons are made inside the CPU and the CPU is also responsible for activating and controlling the operations of other units of a computer system.

Input Unit

- Data and instructions must enter the computer system before any computation can be performed on the supplied data. The input unit that links the external environment with the computer system performs this task. Data and instructions enter input units in forms that depend upon the particular device used. For example, data is entered from a keyboard in a manner similar to typing, and this differs from the way in which data is entered through a mouse, which is another type of input device. However, regardless of the form in which they receive their inputs, all input devices must provide a computer with data that are transformed into the binary codes that the primary memory of the computer is designed to accept. This transformation is accomplished by units that called input interfaces. Input interfaces are designed to match the unique physical or electrical characteristics of input devices to the requirements of the computer system.
- In short, an input unit performs the following functions.
- It accepts (or reads) the list of instructions and data from the outside world.
- It converts these instructions and data in computer acceptable format.
- It supplies the converted instructions and data to the computer system for further processing.

Output Unit

- The job of an output unit is just the reverse of that of an input unit. It supplied information and results of computation to the outside world. Thus it links the computer with the external environment. As computers work with binary code, the results produced are also in the binary form. Hence, before supplying the results to the outside world, it must be converted to human acceptable (readable) form. This task is accomplished by units called output interfaces.
- In short, the following functions are performed by an output unit.
- It accepts the results produced by the computer which are in coded form and hence cannot be easily understood by us.
- It converts these coded results to human acceptable (readable) form.
- It supplied the converted results to the outside world.

Storage Unit

- The data and instructions that are entered into the computer system through input units have to be stored inside the computer before the actual processing starts. Similarly, the results produced by the computer after processing must also be kept somewhere inside the computer system before being passed on to the output units. Moreover, the intermediate results produced by the computer must also be preserved for ongoing processing. The **Storage Unit** or the **primary / main storage** of a computer system is designed to do all these things. It provides space for storing data and instructions, space for intermediate results and also space for the final results.
- In short, the specific functions of the storage unit are to store:

- All the data to be processed and the instruction required for processing (received from input devices).
- Intermediate results of processing.
- Final results of processing before these results are released to an output device.

2.1.2 Software (Operating Systems)

- Software, commonly known as programs, consists of all the electronic instructions that tell the hardware how to perform a task. These instructions come from a software developer in the form that will be accepted by the platform (operating system + CPU) that they are based on. For example, a program that is designed for the Windows operating system will only work for that specific operating system. Compatibility of software will vary as the design of the software and the operating system differ. Software that is designed for Windows XP may experience a compatibility issue when running under Windows 2000 or NT.
- Software is capable of performing many tasks, as opposed to hardware which only perform mechanical tasks that they are designed for. Software is the electronic instructions that tells the computer to perform a task. Practical computer systems divide software systems into two major classes:
  - **System software:** Helps run computer hardware and computer system itself. System software includes operating systems, device drivers, diagnostic tools and more. System software is almost always pre-installed on your computer.
  - **Application software:** Allows users to accomplish one or more tasks. Includes word processing, web browsing and almost any other task for which you might install software. (Some application software is pre-installed on most computer systems.)
- Software is generally created (written) in a high-level programming language, one that is (more or less) readable by people. These high-level instructions are converted into "machine language" instructions, represented in binary code, before the hardware can "run the code". When you install software, it is generally already in this machine language, binary, form

Operating System:

- An operating system is the most important software that runs on a computer. It manages the computer's memory, processes, and all of its software and hardware. It also allows you to communicate with the computer without knowing how to speak the computer's language. Without an operating system, a computer is useless.

  The operating system's job:

  - Your computer's operating system (OS) manages all of the software and hardware on the computer. Most of the time, there are many different computer programs running at the same time, and they all need to access your computer's central processing unit (CPU), memory, and storage. The operating system coordinates all of this to make sure each program gets what it needs.

  Types of operating systems:

  - Operating systems usually come preloaded on any computer you buy. Most people use the operating system that comes with their computer, but it's possible to upgrade or even change operating systems.

- The three most common operating systems for personal computers are Microsoft Windows, Apple Mac OS X, and Linux.

- Modern operating systems use a graphical user interface, or GUI (pronounced gooey). A GUI lets you use your mouse to click icons, buttons, and menus, and everything is clearly displayed on the screen using a combination of graphics and text.

- Each operating system's GUI has a different look and feel, so if you switch to a different operating system it may seem unfamiliar at first. However, modern operating systems are designed to be easy to use, and most of the basic principles are the same.

2.2 Introduction to MS Excel – use of worksheet to enter data, edit data, copy data, move data. Graphical tools in EXCEL for presentation of data.

Introduction:

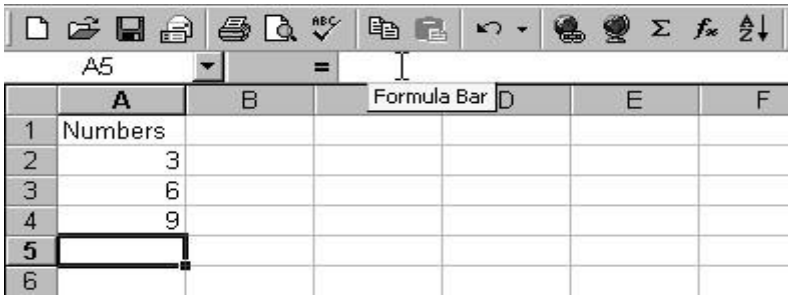
- Microsoft Excel is a spreadsheet developed by Microsoft for Windows, Mac OS X, Android and iOS. It features calculation, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications. It has been a very widely applied spreadsheet for these platforms, especially since version 5 in 1993, and it has replaced Lotus 1-2-3 as the industry standard for spreadsheets. Excel forms part of Microsoft Office.
- Excel used to enter all sorts of data and perform financial, mathematical or statistical calculations.

2.2.2 Use of worksheet to enter, edit/format data, move or copy:

Step 1 - Cell data - Things that can be entered into a cell:

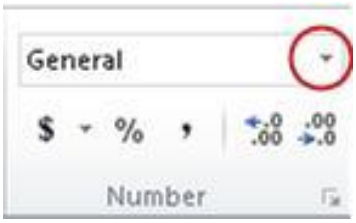
- numbers
- words
- equations, formulas or functions
- fill color
- images (although they are actually on top of a cell, not in it)

Step 2 - Entering data - Move to the cell where you want to enter data and enter words or numbers. If data is already in the cell it will be replaced without you having to cut or delete the previous data.



Step 3 - Format data - Once information has been entered into a cell, you might want to change something about the way the information is displayed. To do that, make sure the cell you want to format is selected and go to the Format menu.

To apply number formatting, click the cell that contains the numbers that you want to format, and then on the Home tab, in the Number group, click the arrow next to General, and then click the format that you want.



To change the font, select the cells that contain the data that you want to format, and then on the Home tab, in the Font group, click the format that you want.

Step-4: Move or copy worksheet data:

- Use the Move or Copy Sheet command to move or copy entire worksheets (also known as sheets), to other locations in the same or a different workbook. Use the Cut and Copy commands to move or copy selected worksheet data to other worksheets or workbooks, or you can drag data between worksheets in different workbooks.

- Although moving or copying the actual worksheet is a fast and effective way to transfer data to another location, you can also move or copy all or part of the data in a worksheet to another worksheet. This method can be used to transfer data to a worksheet in a workbook that is open in a separate instance of Excel.

1. In a worksheet, select the data that you want to move or copy.
2. On the Home tab of the ribbon at the top of the workbook, in the Clipboard group, do one of the following:

- To move the selected data, click Cut  .

  Keyboard shortcut You can also press Ctrl+X.

- To copy the selected data, click Copy  .

Keyboard shortcut You can also press Ctrl+C.

3. Do one of the following:

- Click the worksheet where you want to paste the data.
- Switch to a workbook that is opened in another instance of Excel, and then click the worksheet where you want to paste the data.

4. Select the upper-left cell of the paste area.

Note: Data in the paste area will be overwritten. Also, if the paste area contains hidden rows or columns, you might have to unhide the paste area to see all the copied cells.

5. On the Home tab of the ribbon at the top of the workbook, in the Clipboard group, do one of the following:

Click Paste  .

Keyboard shortcut You can also press Ctrl+V.

Tip: To keep the column width that was originally specified for the data, click the *arrow below Paste*  , click *Paste Special*, and then under Paste, click *Column widths*.

2.2.2 Graphical tools in EXCEL for presentation of data

So many graphical tools are used in EXCEL for presentation of data. The following are the very important graphical tools:

- **Adding a Drop Shadow to a Text Box**
- One way to make your text boxes "stand-off" the page is to add a drop shadow to them. This tip shows just how easy it is to add this formatting touch.

To add a drop shadow to a text box, follow these steps:

1. Make sure the Drawing toolbar is displayed. (You can click on the Drawing tool on the Standard toolbar to display the Drawing toolbar.)
2. Select the text box you want to format. Small selection handles should appear around the perimeter of the text box.
3. Click on the Shadow tool on the Drawing toolbar. Excel displays a palette of available shadows.
4. Click on the shadow desired.

- **Adding Data Labels to Your Chart**
- Adding labels to a chart can make the information presented in the chart more understandable. Excel allows you to add different types of data labels to your charts, as discussed in this tip.

To add data labels, follow these steps:

1. Activate the chart by clicking on it, if necessary.
2. Choose Chart Options from the Chart menu. Excel displays the Chart Options dialog box.

3. Make sure the Data Labels tab is selected. (See Figure) The left side of the dialog box shows the different types of data labels you can choose. (The available types will vary, depending on the type of chart you are using.)

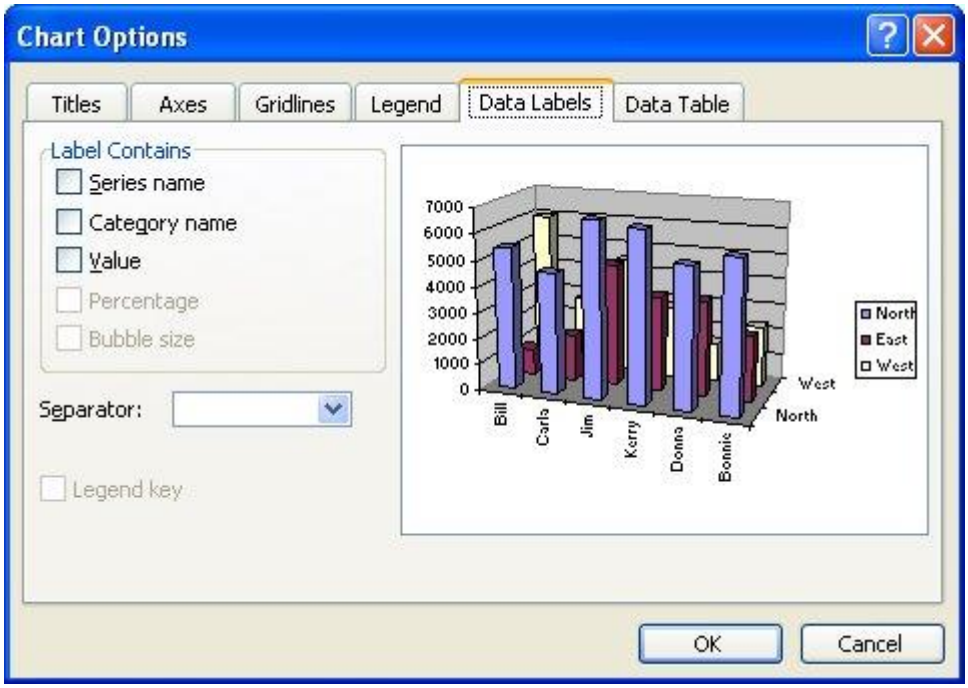


Figure 1. The Data Labels tab of the Chart Options dialog box.

4. There are five different basic types of data labels from which you can choose. Each of them represents a different combination of using the data value, a percentage, or a label as the actual data label. Select the option that best reflects what you want to do. As you make choices, notice that the preview chart is updated according to your selections.
5. Click on OK. Your chart is updated as you directed.

- **Adding Drop Shadows**
- Want your shapes to really "pop" off the page? Add a drop shadow to them, as described in this tip.

To apply a drop shadow to an object,

1. start by selecting the object and then click on the Shadow Style tool. (It is very close to the right end of the Drawing toolbar.)
2. Excel displays a number of different shadow types and positions.
3. You can also modify the shadow, once placed, by choosing the Shadow Settings option.

- **Adding Text to an AutoShape**
- You can add text to all sorts of drawing shapes, not just text boxes. follow these steps:

1. Add your AutoShape as you normally would.
2. Right-click the new AutoShape. Excel displays a Context menu.
3. Choose Add Text from the Context menu. An insertion point appears within the body of the AutoShape.
4. Type your desired text.
5. Click somewhere outside the boundaries of the AutoShape, such as within a cell of the worksheet

- **Changing the Size of a Graphic**
- Adding a graphic to a worksheet is easy. Getting that graphic to just the right size may take a little bit of trial and error. Here's how to adjust the size easily.

Excel allows you to easily resize a graphic you have placed in your workbook by following these steps:

1. Click on the graphic. A box appears around the object (this is designated by eight squares, or frame handles, around the outside of the graphic).
2. Use the mouse to point to one of the frame handles. Click on the left mouse button.
3. Drag the frame handle to resize the graphic.
4. Release the mouse button when the graphic is the size you want.

- **Colorizing Charts**
- Need to change the color of different parts of your chart? It's easy to do when you apply the technique described in this tip.

For pie charts, follow these steps:

1. Click on the "pie" so that it is surrounded by handles (little squares).
2. Click again on the section you want to change. The handles will now surround only that section.
3. Right-click on the section. Excel displays a Context menu.
4. Choose the Format Data Point option from the Context menu. Excel displays the Format Data Point dialog box, with the Patterns tab selected. (See Figure)

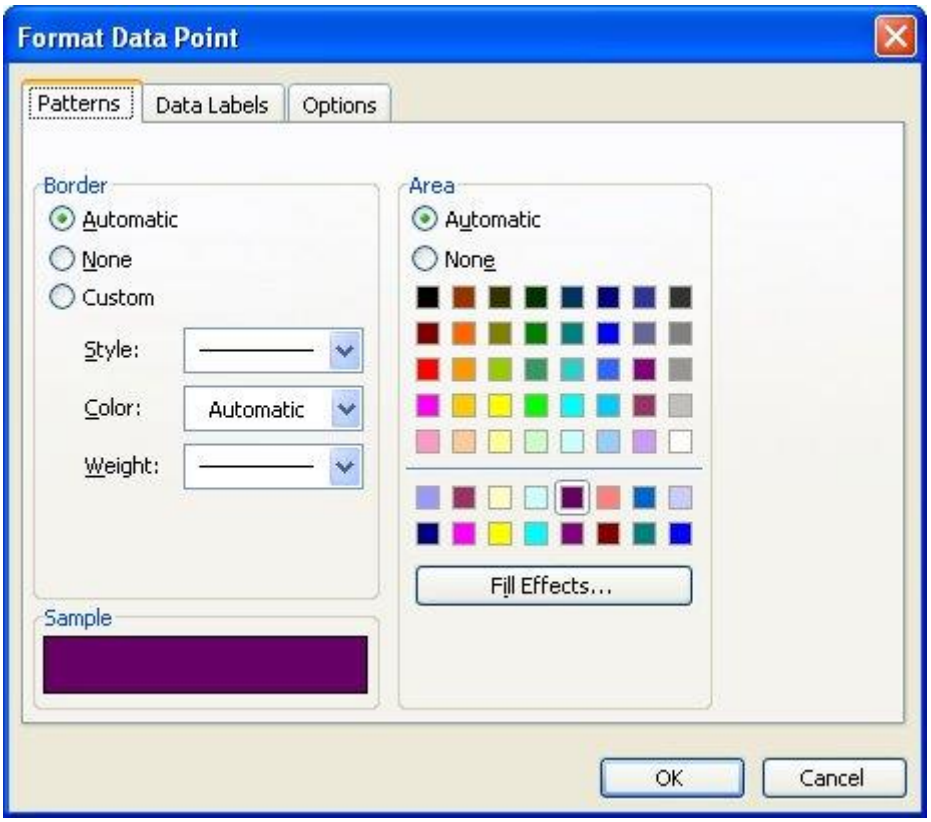


Figure . The Format Data Point dialog box.

5. In the Area portion of the dialog box, select the color you want to use for the chart section.
6. Click on OK. Excel updates your chart.

These steps can be easily adapted to any type of chart. The only difference is that you select the chart object (bar, point, what have you) in the first two steps instead of the pie section.

- **Creating a Drawing Object**
- Creating simple drawing objects is easy in Excel. All you need to do is use the tools made available on the Drawing toolbar.

On the left side of the Drawing toolbar are several tools that are used to create basic shapes. The line, arrow, rectangle, and oval tools are easily identifiable. In addition, you can use the AutoShapes tool to display a menu of more than 125 different symbols, banners, and callouts. Follow these steps to place a drawing object in your worksheet:

1. Make sure the Drawing toolbar is displayed.
2. Click on the tool that represents the type of object you want to create. If you are creating an AutoShape, click on the AutoShape tool, then choose the shape from the appropriate submenu. Once a tool is selected, the mouse pointer changes to crosshairs, or a plus sign.
3. Click within your worksheet at one corner of where you want the shape to appear.
4. Drag the mouse to the opposite corner for the object.
5. When you release the mouse button, the object appears in the worksheet and you can manipulate it as desired.

- **Using Chart Titles**
- Titles can be a great addition to any chart. They help provide explanatory information about the information in the chart. Here's the quick way to add all the titles you need.

You should note that the titles available for any given chart will vary, depending on the type of chart you are using. For instance, the only type of title available with a pie chart is the chart title itself. Since there are no X, Y, and Z axes on a pie chart, there are no titles available for them.
To insert titles, follow these steps:

1. Activate the chart by clicking on it, if necessary.
2. Choose Chart Options from the Chart menu. Excel displays the Chart Options dialog box.
3. Make sure the Titles tab is selected. (See Figure)

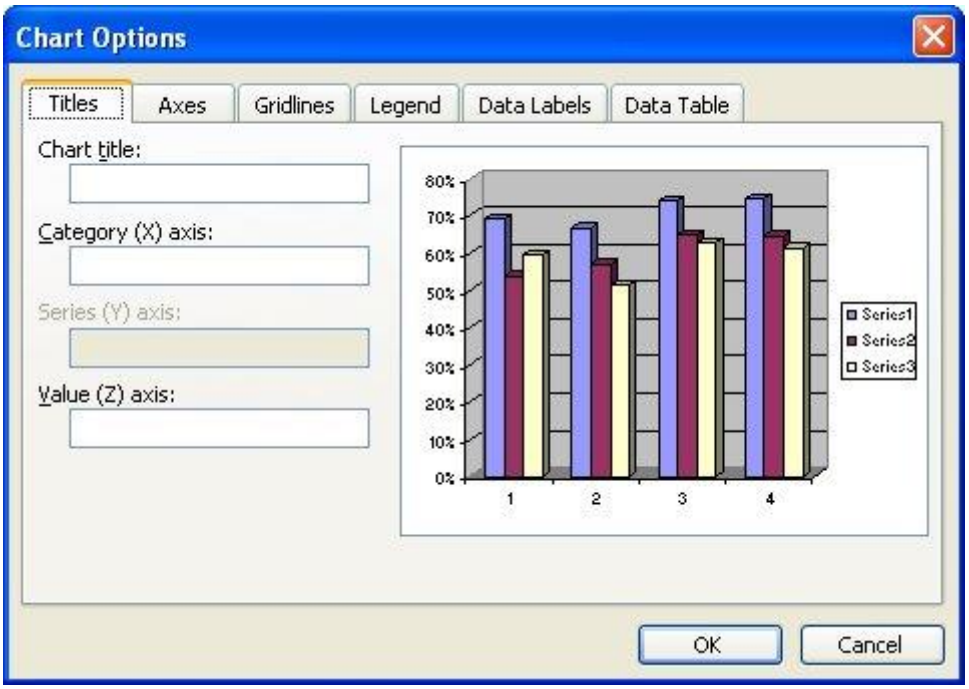


Figure. The Titles tab of the Chart Options dialog box.

4. Fill in the appropriate fields in the dialog box for the titles you wish to add. If you leave a field blank, the title will not be included.
5. When finished, click on OK. The titles are updated as you directed.

To change the text contained within a title, you can always follow these same steps again. There is an easier, more direct way to change title text, however. All you need to do is click on the title. The title is selected and surrounded with a box and handles. When you move the mouse pointer inside the box, it changes to an insertion point. Click the mouse pointer once to edit the text.

Notice that if you press **Enter**, the cursor only moves to the next line; you are still adding to the title. To signify that you are finished entering title text, you must use the mouse pointer to select some other part of your chart or worksheet.

When you add titles to your chart, Excel places them in a position it feels is best for the title. Thus, an axis title will be centered on the axis. You can move titles very easily, however. To do this, use the mouse to select the title text. When you do, it becomes surrounded with a box and handles. Use the mouse to point to the border around the title. Press and hold down the mouse button. As you move the mouse, the title is also moved. When you release the mouse button, the title remains at the new location.

2.3 MS-WORD – editing, copying, moving, formatting, table insertion, drawing flow charts etc.,

---

- **Microsoft Word** is a word processor developed by Microsoft. It was first released on October 25, 1983[3] under the name *Multi-Tool Word* for Xenix systems. Subsequent versions were later written for several other platforms including IBM PCs running DOS (1983), Apple Macintosh running Mac OS (1985), AT&T Unix PC (1985), Atari ST (1988), OS/2 (1989), Microsoft Windows (1989) and SCO Unix (1994). Commercial versions of Word are licensed as a standalone product or as a component of Microsoft Office, Windows RT or the discontinued Microsoft Works suite. Microsoft Word Viewer and Office Online are Freeware editions of Word with limited features.

Editing Document

Typing and inserting Text

To enter text, just start typing! The text will appear where the blinking cursor is located. Move the cursor by using the arrow buttons on the keyboard or positioning the mouse and clicking the left button. The keyboard shortcuts listed below are also helpful while moving through the text of a document:

Move Action	**Keystroke**
Beginning of the line	**HOME**
End of the line	**END**
Top of the document	**CTRL+HOME**
End of the document	**CTRL+END**

Selecting Text

To change any attributes of text it must be highlighted first. Select the text by dragging the mouse over the desired text while keeping the left mouse button depressed, or hold down the

SHIFT key on the keyboard while using the arrow buttons to highlight the text. The following table contains shortcuts for selecting a portion of the text:

Selection	Technique
Whole word	double-click within the word
Whole paragraph	triple-click within the paragraph
Several words or lines	drag the mouse over the words, or hold down **SHIFT** while using the arrow keys
Entire document	choose **Editing** | **Select** | **Select All** from the Ribbon, or press **CTRL+A**

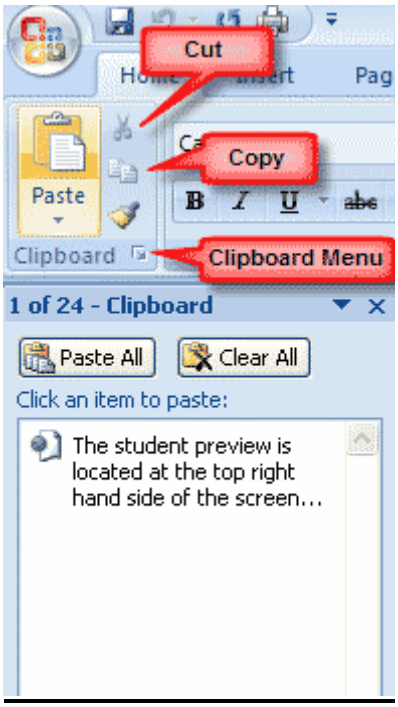
Deselect the text by clicking anywhere outside of the selection on the page or press an arrow key on the keyboard.

Inserting Additional Text

Text can be inserted in a document at any point using any of the following methods:

- **Type Text**: Put your cursor where you want to add the text and begin typing
- **Copy and Paste Text**: Highlight the text you wish to copy and right click and click **Copy**, put your cursor where you want the text in the document and right click and click **Paste**.
- **Cut and Paste Text**: Highlight the text you wish to copy and right click and click **Cut**, put your cursor where you want the text in the document and right click and click **Paste**.
- **Drag Text**: Highlight the text you wish to move, click on it and drag it to the place where you want the text in the document.

You will notice that you can also use the Clipboard group on the Ribbon.



To copy a comment, follow these steps:

1. Highlight the comment mark in your document.
2. Press **Ctrl+C**. The comment mark and the associated comment are copied to the Clipboard.

3. Position the insertion point at the location where you want to copy the comment.
4. Press **Ctrl+V**. The comment mark is inserted in your document, and the associated comment is added to your document.

To move a comment to another location in your document (or even to another document), you can use techniques you already know for moving regular text. To move a comment, follow these steps:

1. Highlight the comment mark for the comment you want to move.
2. Press **Ctrl+X**. The comment mark and the associated comment are removed from your document and copied to the Clipboard.
3. Position the insertion point at the location where you want the comment moved.
4. Press **Ctrl+V**. The comment mark is inserted in your document, and the associated comment is again added to the document.

To remove a comment from your document, follow these steps:

1. Highlight the comment mark in your document.
2. Press either **Del** or **Ctrl+X**.

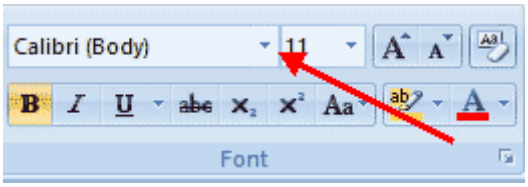
Formatting :

Styles : A style is a format enhancing tool that includes font typefaces, font size, effects (bold, italics, underline, etc.), colors and more. You will notice that on the Home Tab of the Ribbon, that you have several areas that will control the style of your document - Font, Paragraph, and Styles.

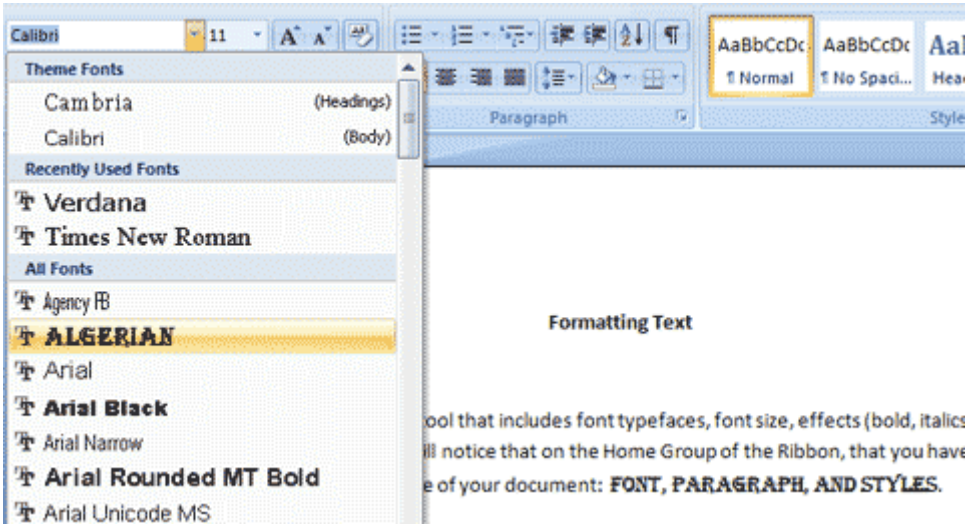
Change Font Typeface and Size

To change the font typeface:

- Click the **arrow** next to the font name and choose a font.

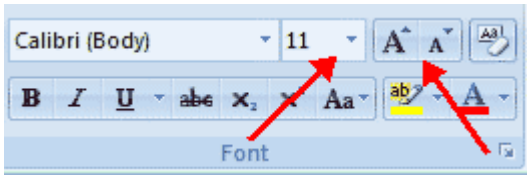


- Remember that you can preview how the new font will look by highlighting the text, and hovering over the new font typeface.



To change the font size:

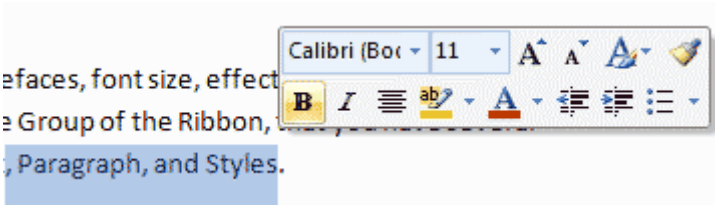
- Click the **arrow** next to the font size and choose the appropriate size, or
- Click the **increase** or **decrease** font size buttons.



Font Styles and Effects

Font styles are predefined formatting options that are used to emphasize text. They include: Bold, Italic, and Underline. To add these to text:

- Select the text and click the **Font Styles** included on the Font Group of the Ribbon, or
- Select the text and right click to display the font tools



Change Text Color

To change the text color:

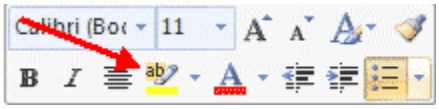
- Select the text and click the **Colors** button included on the Font Group of the Ribbon, or
- Highlight the text and right click and choose the colors tool.
- Select the color by clicking the down arrow next to the font color button.

Highlight Text :

Highlighting text allows you to use emphasize text as you would if you had a marker. To highlight text:

- Select the text
- Click the **Highlight Button** on the Font Group of the Ribbon, or

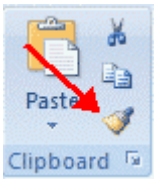
- Select the text and right click and select the highlight tool
- To change the color of the highlighter click on down arrow next to the highlight button.



Copy Formatting

If you have already formatted text the way you want it and would like another portion of the document to have the same formatting, you can copy the formatting. To copy the formatting, do the following:

- Select the text with the formatting you want to copy.
- Copy the format of the text selected by clicking the **Format Painter** button on the Clipboard Group of the Home Tab
- Apply the copied format by selecting the text and clicking on it.



Clear Formatting

To clear text formatting:

- Select the text you wish to clear the formatting
- Click the **Styles** dialogue box on the Styles Group on the Home Tab
- Click **Clear All**

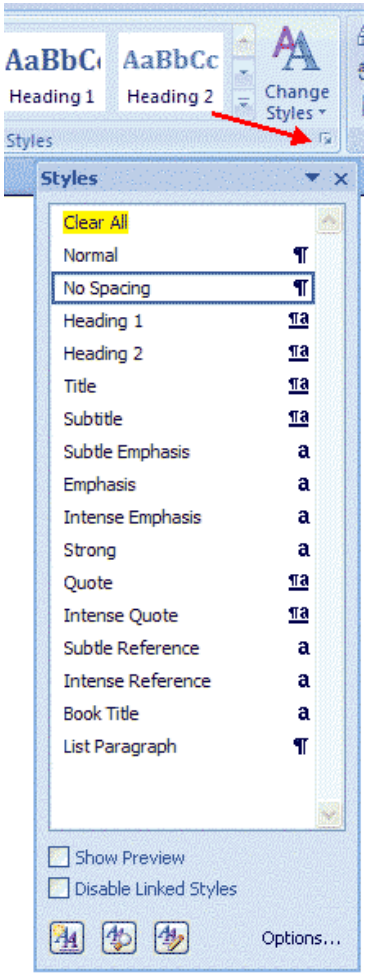
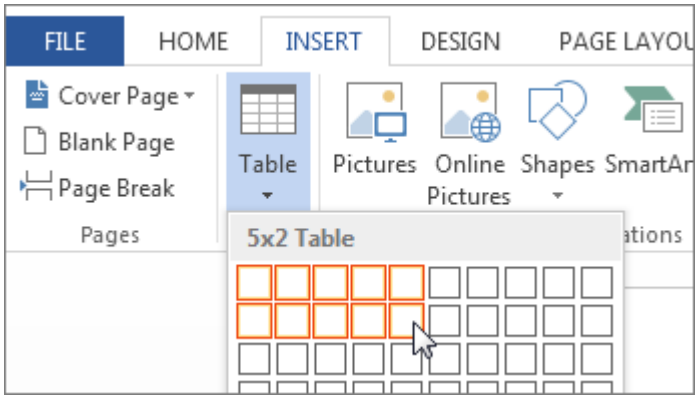


Table Insertion:

Insert a table

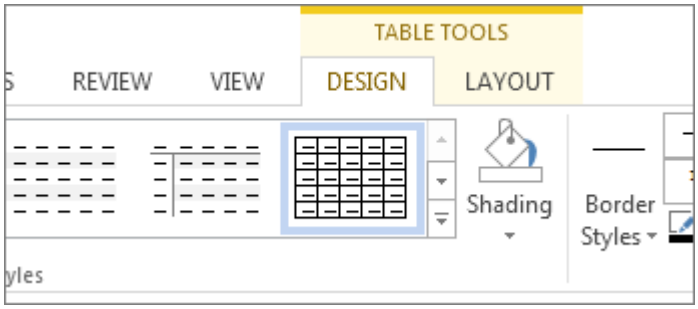
In Microsoft Office Word 2007, you can insert a table by choosing from a selection of preformatted tables — complete with sample data — or by selecting the number of rows and columns that you want. You can insert a table into a document, or you can insert one table into another table to create a more complex table.

To quickly insert a basic table, click **Insert** > **Table** and move the cursor over the grid until you highlight the number of columns and rows you want.



Click and the table appears in the document. If you need to make adjustments, you can add table rows and columns, delete table rows and columns, or merge table cells into one cell.

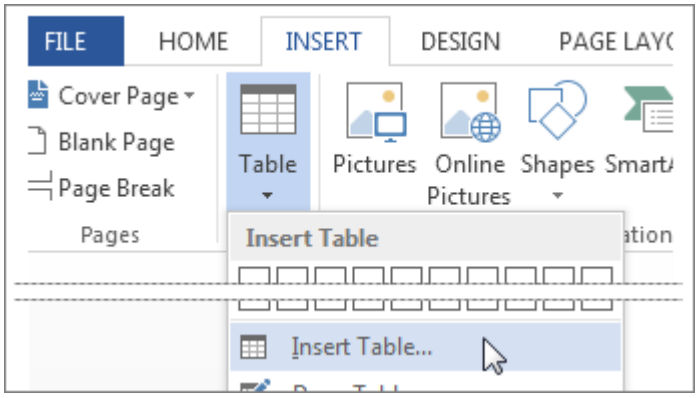
When you click in the table, the **Table Tools** appear.



Use **Table Tools** to choose different colors, table styles, add a border to a table or remove borders from a table. You can even insert a formula to provide the sum for a column or row of numbers in a table.

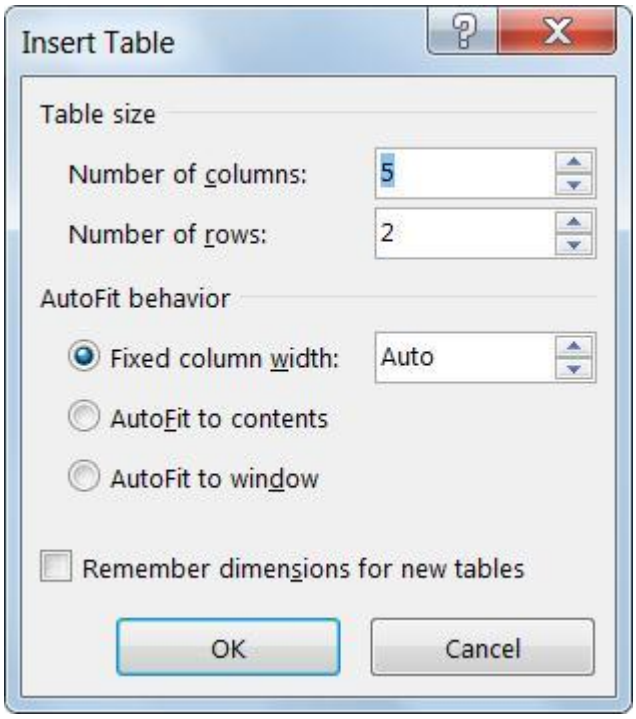
Insert larger tables or tables with custom width behaviors

For larger tables and for more control over the columns, use the **Insert Table** command.



This way you can create a table with more than ten columns and eight rows, as well as set the column width behavior.

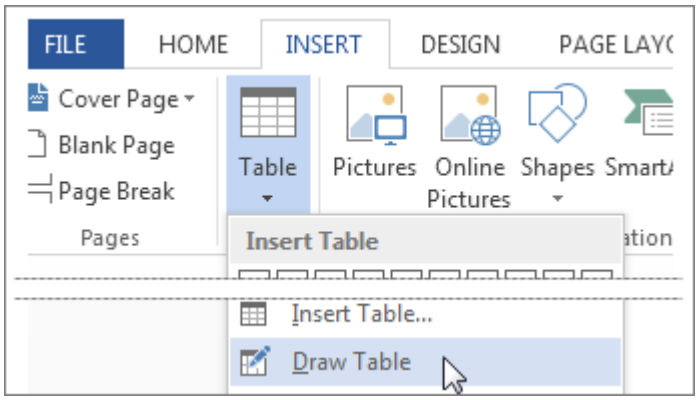
1. Click **Insert** > **Table** > **Insert Table**.
2. Set the number of columns and rows



3. In the **AutoFit behavior** section you have three options for setting how wide your columns are:
    - **Fixed column width**: You can let Word automatically set the column width with Auto, or you can set a specific width for all of your columns.
    - **AutoFit to contents**: This will create very narrow columns that will expand as you add content.
    - **AutoFit to window**: This automatically changes the width of the entire table to fit the size of your document.
4. If you want each table you create to look like the table you're creating, check **Remember dimensions for new tables**.

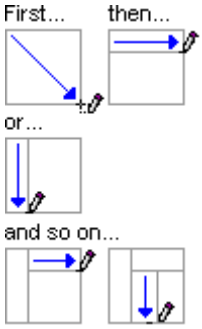
Design your own table

If you want more control over the shape of your table's columns and rows or something other than a basic grid, the **Draw Table** tool helps you draw exactly what you want.



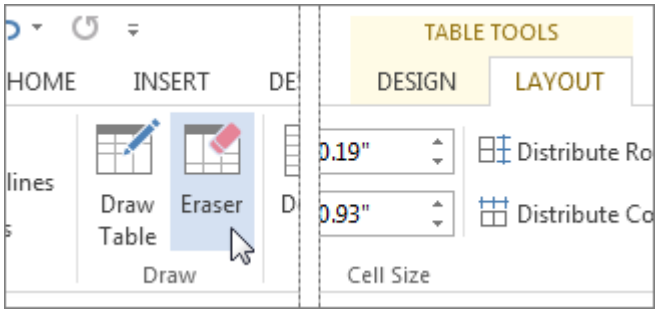
You can even draw diagonal lines and cells within cells.

1. Click **Insert** > **Table** > **Draw Table**. The pointer changes to a pencil.
2. Draw a rectangle to make the table's borders. Then draw lines for columns and rows inside the rectangle.



3. To erase a line, click the **Table Tools Layout** tab, click **Eraser**, and then click the line that you want to erase.

4.



2.4 Introduction to PPT, image, data handling and Graphical tools in PPT for presentation

2.4.1 Introduction to PPT:

Microsoft PowerPoint is a software product used to perform computer-based presentations. There are various circumstances in which a presentation is made: teaching a class, introducing a product to sell, explaining an organizational structure, etc.

Microsoft PowerPoint, part of Microsoft Office, creates and plays presentations. A presentation is something a speaker makes to an audience, typically using a computer and LCD projector to display material in a lecture hall or auditorium. PowerPoint works a lot like Microsoft Word, and the assumption here is that you are familiar with Word.

A PowerPoint presentation is made up of "slides" that are individual frames or screens of information. To create a presentation, create the slides. A PowerPoint file (*.ppt) is a collection of slides, typically for one and only one presentation, although files can be linked together to make up compound presentations. PowerPoint has functions for:

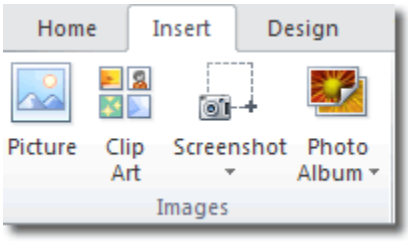
- Creating and inserting new slides.
- Editing existing slides.
- Reordering existing slides.

The main purpose of MS PowerPoint is to enable the user to create dynamic, informational slide shows through the use of text, graphics, and animation. Slide shows created with the software are often displayed on projection screens for business, training,

or educational presentations, although they can be distributed as stand-alone files. Additionally, the slides can be arranged and printed as handouts for reference.

2.4.2 Image, data handling and Graphical tools in PPT for presentation

To insert an image in PowerPoint, click Insert > Images. You can see hat there are four different types of image you can insert here, and we'll explore each one now.

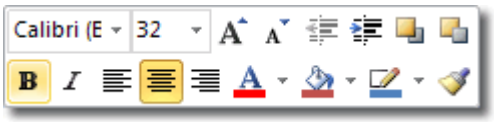


Insert A Picture

If you have an image, for example a photo you have taken, on your PC's hard drive, you would click on the Picture button. Then you would navigate to the place on your hard drive where the picture was located and either double click on it or select it and then click Insert.

Data handling:

To format some text, first of all select it. As you move the cursor, the mini toolbar appears as if by magic. The mini toolbar contains some of the more commonly used formatting commands that PowerPoint gueses you are likely to use. Using it, you can bold text, italicize it and do various other text formatting tasks, all at the click of a button. You will probably be familiar with all of the commands available on the mini toolbar, so we won't dwell on them.



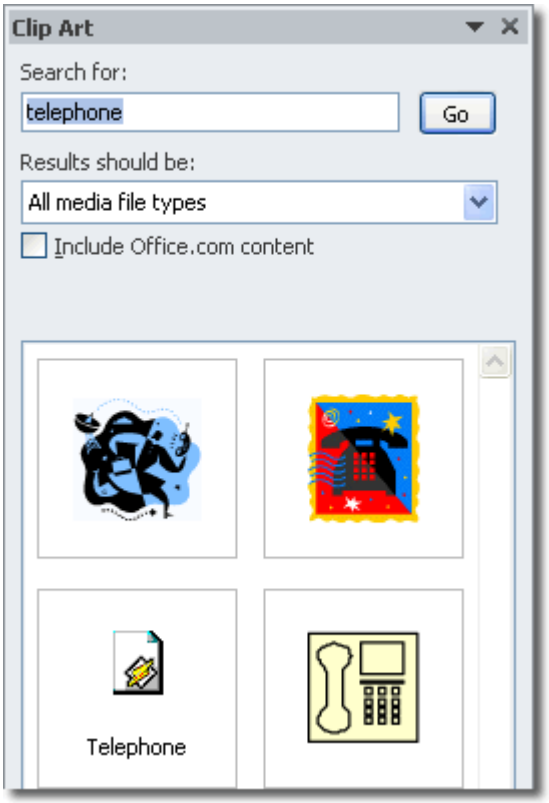
If you want to apply more adventurous formatting, head over to the ribbon. When you select text, the Format contextual tab appears. On this tab are many different formatting functions. Let's get to grips with formatting text by running through a quick example.

Let's increase the font size and bold some text. To do that, all we have to do is select the text and then press ctrl-b. To increase the font size incrementally, press ctrl-shift->. You can keep pressing it until you get the right size, or select a specific font size from the ribbon. Click on the Format contextual tab and then click on the More button in WordArt styles.

Graphical Tools:

a) Using clip art

PowerPoint provides a large selection of ready made Clip Art images. These images are simple in design but their use can really help get your point across in your presentation. One advantage of using Clip Art is that the collection of images is right there raedy for you to use. You don't have to go out and take photos for your presentation, and you don't have to touch Photoshop. When you click the Clip Art button, the Clip Art panel opens on the right of the workspace.



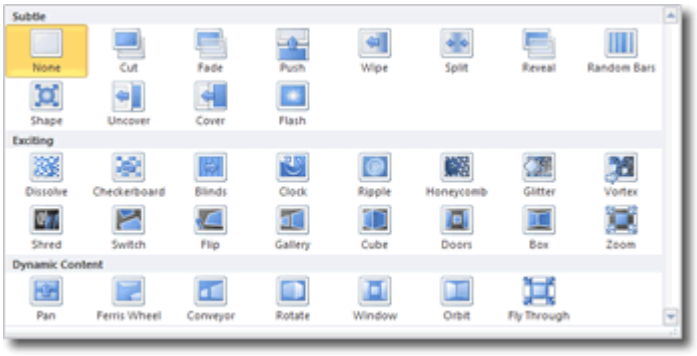
We can use this panel to search for Clip Art images of certain things. In the image above, I searched for "telephone" and found four related images. To narrow down the saerch, you can click on the Results should be drop down list and select one of the categories of:

Illustrations
Photographs
Videos
Audios

To insert a Clip Art image, click on it. You'll notice that the Picture Tools contextual tab appears in the ribbon to help you perform picture related tasks. This tab will remain visible as long as the item of Clip Art (or, indeed, any image) is selected.

b) Adding Transitions:

In Microsoft PowerPoint, slide transitions are motion effects that occur in Slide Show view when you move from one slide to the next during a presentation. You can control the speed, add sound, and even customize the properties of transition effects. To add a transition first of all select a slide in the left hand panel that contains the Slides and Outline tabs (I'd keep it simple and work in the Slides tab). The transition will occur immediately before the selected slide is displayed. Click Transitions > Transition To This Slide, and then click on the transition you want to use. If you want to see a larger selection of transitions, click on the more button (the down arrow at the bottom right of the group.

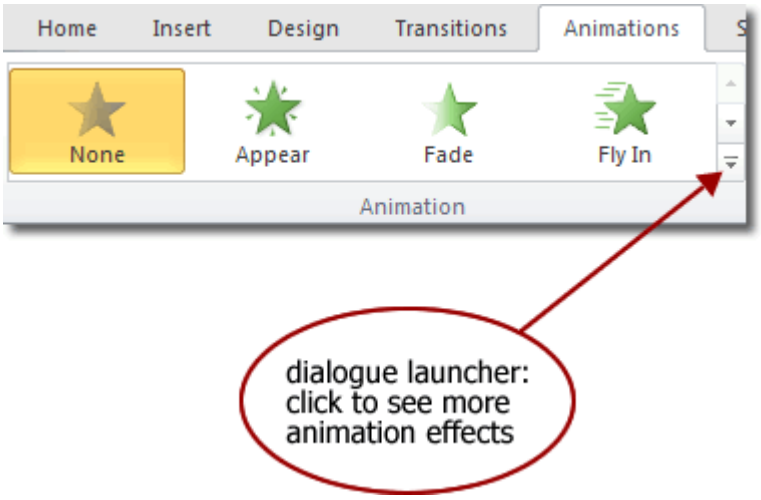


When you hover over a thumbnail image of a transition, you can see a live preview of it applied to your slide. When you move the cursor away, the preview is removed.

If, after having applied a transition to the selected slide, you decide that you want the transition applied to all slides, click on the Apply To All button in the Timing Group on the same tab.

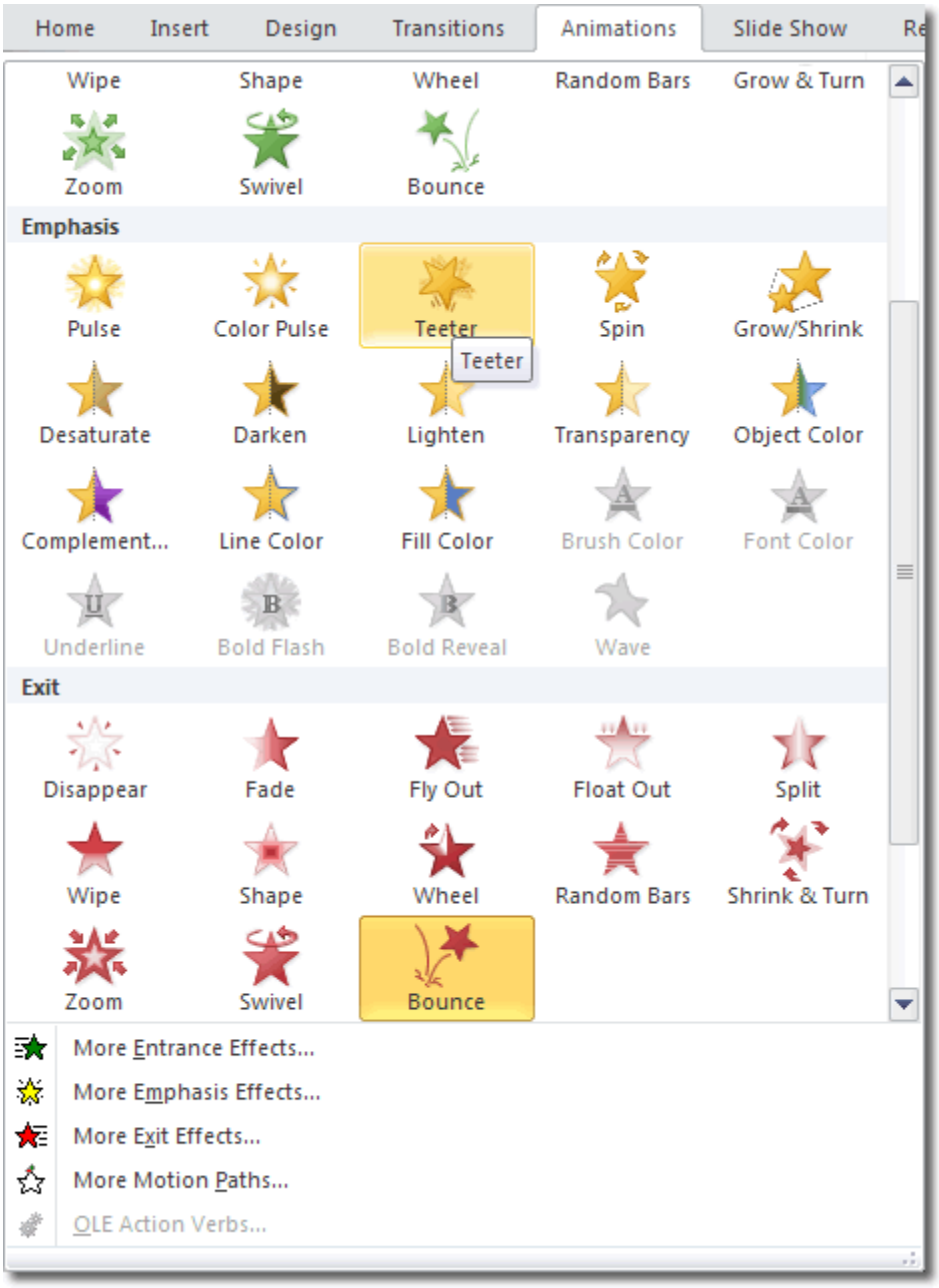
c) Creating Animations:

To add a custom animation to an object in Microsoft PowerPoint, first of all select the object, then click animation in the ribbon to see all the animation options we have at our disposal. We can see a few thumbnail images representing entrance effects in the animation group.



That's not a very big selection of effects. To see more effects, click on the dialogue launcher at the bottom right of the Animation group. A more comprehensive selection of effects is then displayed in a thumbnail gallery, in the categories of:

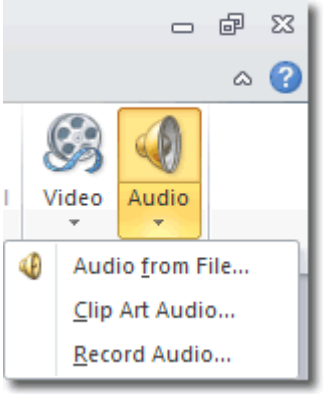
- Entrance - this kind of animation brings the object onto the PowerPoint slide from a location off the slide
- Emphasis - once the object is on the slide, you can give it some emphasis with this kind of animation
- Exit - animations in this category remove the object from a slide
- Motion Paths - applying this kind of animation allows us to plot a path for the object to follow



d) Adding Sounds:

There are a few different ways to add a sound clip in PowerPoint.

To insert a sound clip into your PowerPoint presentation, click Insert > Media > Ausio. You'll notice that the audio button is split into two. Click on the lower half of the button and you'll find the following options available to you:

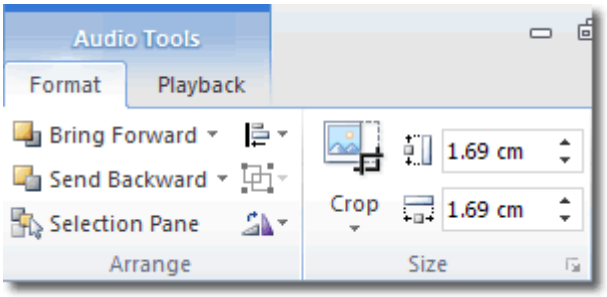


We'll go through each option now.

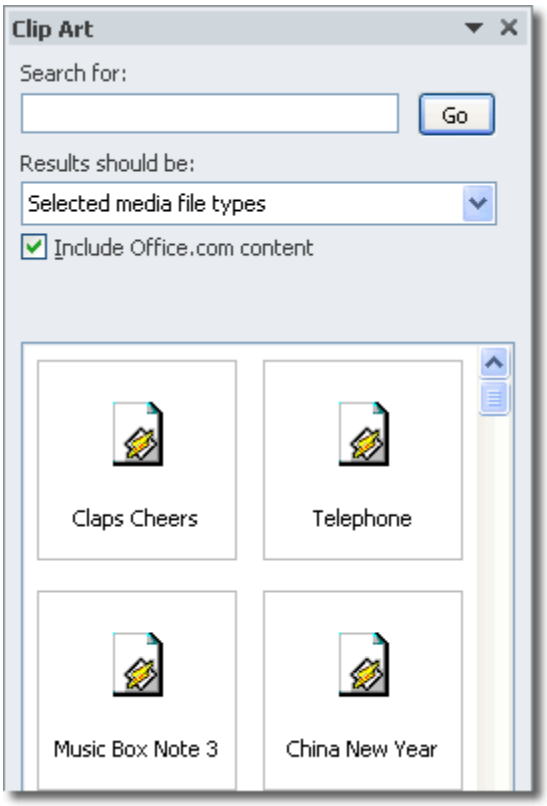
Audio From File: This is the most straight forward option. If you have a sound file stored somewhere on your computer, use this option to navigate to it, select it and insert it. The icon representing a sound clip is inserted onto the slide you had selected.

While the sound clip is selected, you will be able to see the basic sound playback tools below the clip: play, rewind, fast forward and volume controls. You can also position the playback point to a specific place in playback using the timeline. The progress in minutes and seconds also appears to the right of the timeline, to let you know how far through the recording you've progressed.

While the audio clip is selected, you'll also see the Adio Tools contextual tab displayed in the ribbon, allowing you to perform a variety of audio tasks on the clip.



The Format tab mostly contains commands for use with video, but the Playback tab can be used with sound clips.



Clip Art Audio

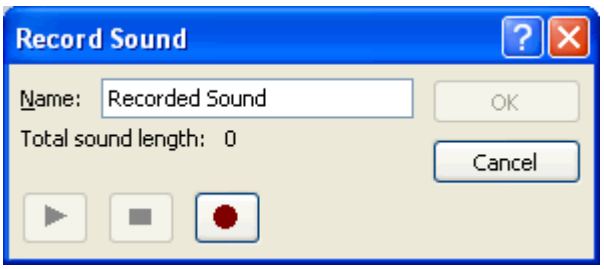
You can use the Clip Art library to insert pre made audio clips in much the same way that you would insert Clip Art images. When you click The Clip Art Audio option, the Clip Art panel appears to the right of the work space but the Results should be selection list is set to Audio.

Type in the type of sound you need into the search box and press Enter or click Go. A selection of sound clips will be displayed for you to insert. Click on a particular clip and it will be inserted.

Record Audio

The record audio option gives you the ability to record your own sound clip within PowerPoint. All you need is a microphone to record your voice.

When you take this option, PowerPoint displays a dialogue box that you can use to start and stop recording your voice. Type the name of the clip in the Name box and then click the round and red record button on the right. Stop the recording by clicking the blue rectangle button in the centre and start playback using the blue triangle button on the left.



2.5 Use of Computers in Data Processing and Mapping.

Computer:
A computer is an electronic device. It consists of various sub-systems such as memory, micro-processor, input system (keyboard) and output system (Printer). It is an extremely powerful device. A computer is a fast and versatile machine that can perform simple to complex functions without intervention by a human operator during the run.

Advantages of using computer:
1. It increases the speed of the computation and data processing.
2. It can handle huge volume of the data, which is normally not possible manually.
3. It facilitates copy, edit, save and retrieve the data at will.
4. It further enables validation, checking and correction of data easily.
5. Computer makes it very easy to perform comparative analysis, whether by drawing maps or graphs.
6. The type of graph or map (i.e. bar/pie or types of shades), heading, indexing and other formats can be changed very easily.

Hardware:
The hardware components of a computer include:
(a) A Central Processing Unit (CPU) and Storage System
(b) A Graphic Display Sub-system
(c) Input Devices
(d) Output Devices

1. A Central Processing Unit and Storage System
- The CPU is the 'mind' of modern computers. The processor in the CPU executes and processes data and controls other equipments. Hard disk is used for data storage. The Random Access Memory (RAM), the secondary storage such as floppy disks, CD ROM, pen drives, and magnetic tapes are parts of CPU and used for data storage.

2. A Graphic Display System or Monitor
- A graphic display system or monitor serves as the visual communication medium in all computers.

3. Input Devices

- The instruction and the data are entered into the computer using the keyboard. Scanners and digitisers of different size and capabilities are also used for spatial data entry.

4. Output Devices

- The output devices include a variety of printers such as ink-jet, laser and colour laser printers; and the plotters that are available in different sizes ranging from A3 to A0 size.

Computer Software

Software is the written program made up of electronic codes and is stored in memory. It performs specific functions as per the instructions given by the user. Operating system forms base of computer such as Windows and Linux. Work software such as MS Excel/Spread sheet, Lotus 1 – 2 – 3, and d – base, Openoffice Math. Arc View/Arc GIS, Geomedia.

Using MS Excel or Spreadsheet for data processing:

- MS Excel, Lotus 1 – 2 – 3, and d – base are some of the important softwares used for data processing, and drawing graphs and diagrams. MS Excel being most widely used and commonly available software program and it is also compatible with map-making software as one can easily feed data in MS Excel and attach it to the map-making software to create maps.
- MS Excel displays the worksheet, which consists of rows and columns. The intersection of a row and column is a rectangular area, which is called a cell. In other words, a worksheet is made up of cells. A cell can contain a numerical value, a formula (which after calculation provides numerical value) or text.
- An Excel worksheet contains 16,384 rows, numbered 1 through 16384 and 256 columns, represented by default through letters A through Z, AA through AZ, BA through BZ, and continuing to IA through IZ. By default, an Excel workbook consists of three worksheets. If you require, you can insert more, up to 256 worksheets.

S. No.	Function	Instructions	Menu	Secondary Menu (from dropdown list)	Keyboard Shortcuts
1.	For opening a new file		File	New	Ctrl N
	For opening an existing file		File	Open	Ctrl O
2.	Save a file	Give a file name and define where you want to store it (by default, it is c:\....\my documents\)	File	Save	Ctrl S
3.	Copy, move and paste a set of data	Select the set of data by pressing the left mouse button and dragging it over the set of the data you want to select	Edit	Copy	Ctrl C
4.	Cut, move and paste a set of data	Select the set of data by pressing the left mouse button and dragging it over the set of the data you want to select	Edit	Cut	Ctrl X
5.	Paste a set of data	Take the cursor to the cell where you want to paste it	Edit	Paste	Ctrl V
6.	For undoing the last action*		Edit	Undo	Ctrl Z
7.	For redoing the last action*		Edit	Repeat	Ctrl Y

Example 1: Data analysis in Excel

Solving the expression "5 + 6 – 8 – 5"

Step 1: Click on any cell (with the help of mouse).
Step 2: Type =, followed by the expression.
Thus, the expression becomes = 5 + 6 – 8 – 5.
Step 3: Press enter key, and you will get the result in the same cell that you had chosen in Step 1.
Note: The numerical operations can only be performed in excel by first typing = sign.

Example 2: Calculating percentage in Excel

Step 1: Enter the name of the states in first column (i.e. column A).

Step 2: In Column B, corresponding to each state, enter the size of urban population.
Step 3: In Column C, corresponding to respective state enter the size of total Population.
Step 4: In Column D and row 2, type = followed by B2/C2 (that is total urban population of Andhra Pradesh divided by the total population in the same State) and *100 (multiplied by 100). Thus, the expression becomes =B2/C2*100.
Step 5: Press enter key. This will give you solution of the expression, that is, the percentage of urban population in Andhra Pradesh.
Step 6: Now you need not to write the formula again for calculating percentage of urban population for other states. Simply, click on the cell D2. This will copy the formula of the first state/cell to all the downward cells you have dragged it over.

(Note: the formula =B2/C2*100 that has been written in cell D2, and becomes B3/C3*100 in cell D3, and so on).

Computer Assisted Mapping

The maps may also be drawn using a combination of computer hardware and the mapping software. The computer assisted mapping essentially requires the creation of a spatial database and non – spatial data.

1. Spatial Data
The spatial data represent a geographical space. They are shown by the points, lines and the polygons. To show the location of schools, hospitals, wells, tube-wells, towns and villages, etc. on the map we use points. Similarly, lines are used to show linear features like roads, railway lines, canals, rivers, power and communication lines, etc. Polygons are used to show area features such as administrative units (countries, districts, states, and blocks); land use types (cultivated area, forest lands, degraded/waste lands, pastures, etc.) and features like ponds, lakes, etc.

2. Non–Spatial Data
The data describing the information about spatial data are called as non-spatial data. For example, we can attach the information such as the name, number, facilities, etc.

Mapping Software and their Functions

There are a number of commercially available mapping softwares such as ArcGIS, ArcView, Geomedia, GRAM, Idrisi, Geometica, etc. Mapping software provides functions for spatial and attribute data input. It helps in digitisation of scanned maps, corrections of errors, transformation of scale and projection, data integration, map design, presentation and analysis.

* * * * *

Unit-III Internet Basics

3.1 Introduction to internet – Basics and Applications of Internet, Internet working, Internet access.
3.2 Using of Internet – understanding the basics, title , menu and tool bars, Address bar, Navigating web pages and Web sites and Printing.
3.3 Understanding the World Wide Web (WWW)
3.4 Searching Tools – World Search Engines, Search Directories and encyclopedias.
3.5 Online safety – Spywares and viruses.

3.1 Introduction to internet – Basics and Applications of Internet, Internet working, Internet access

3.1.1 Basics and Applications of Internet:

Introduction:

- The Internet is a worldwide telecommunications system that provides connectivity for millions of other, smaller networks; therefore, the Internet is often referred to as a network of networks. It allows computer users to communicate with each other across distance and computer platforms.

- The Internet began in 1969 as the U.S. Department of Defense's Advanced Research Project Agency (ARPA) to provide immediate communication within the Department in case of war. Computers were then installed at U.S. universities with defense related projects. As scholars began to go online, this network changed from military use to scientific use. As ARPA net grew, administration of the system became distributed to a number of organizations, including the National Science Foundation (NSF). This shift of responsibility began the transformation of the science oriented ARPAnet into the commercially minded and funded Internet used by millions today.

- The Internet acts as a pipeline to transport electronic messages from one network to another network. At the heart of most networks is a server, a fast computer with large amounts of memory and storage space. The server controls the communication of information between the devices attached to a network, such as computers, printers, or other servers.

- An Internet Service Provider (ISP) allows the user access to the Internet through their server. Many teachers use a connection through a local university as their ISP because it is free. Other ISPs, such as America Online, telephone companies, or cable companies provide Internet access for their members.

- The picture below illustrates two computers connected to the Internet; your computer with IP address 1.2.3.4 and another computer with IP address 5.6.7.8. The Internet is represented as an abstract object in-between. (As this paper progresses, the Internet portion of Diagram 1 will be explained and redrawn several times as the details of the Internet are exposed.)

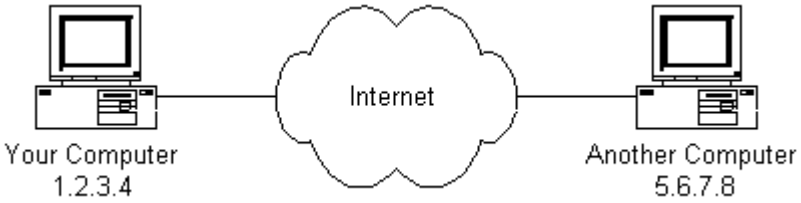


Diagram 1

- You can connect to the Internet through telephone lines, cable modems, cellphones and other mobile devices. If you connect to the Internet through an Internet Service Provider (ISP), you are usually assigned a temporary IP address for the duration of your dial-in session. If you connect to the Internet from a local area network (LAN) your computer might have a permanent IP address or it might obtain a temporary one from a DHCP (Dynamic Host Configuration Protocol) server. In any case, if you are connected to the Internet, your computer has a unique IP address.

3.1.2 Applications of Internet:

Some of the important services provided by Internet are:

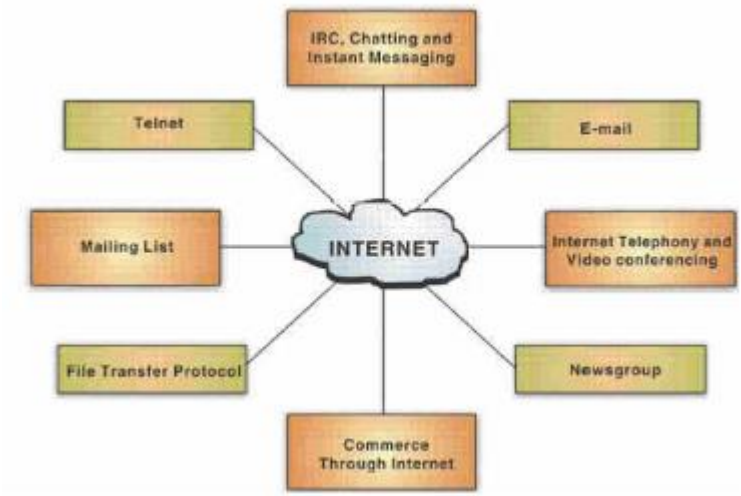
World Wide Web (WWW): It is a subset of the Internet and it presents text, images, animation, video, sound, and other multimedia in a single interface. The operation of the Web relies primarily on hypertext, as it is a means of information retrieval.

Electronic Mail (E-Mail): It is the process of exchanging messages electronically, via a communications network, using the computer.

File Transfer Protocol (FTP): It is a system of rules and a software program that enables a user to log on to another computer and transfer information between it and his/her computer.

Telnet: It connects one machine to another in such a way that a person may interact with another machine as if it is being used locally.

Internet Relay Chat (IRC): This service allows people to communicate in real time and carry on conversations via the computer with one or more people. It provides the user with the facility to engage in simultaneous (synchronous) online 'conversations' with other users from anywhere in the world.



Chatting and Instant Messaging: Chat programs allow users on the Internet to communicate with each other by typing in real time. Instant messaging allows a user on the Web to contact another user currently logged in and type a conversation. Internet Telephony: It refers to the use of the Internet rather than the traditional telephone company infrastructure, to exchange spoken or other telephonic information.

Video Conferencing: It uses the same technology as IRC, but also provides sound and video pictures. It enables direct face-to-face communication across networks via web cameras, microphones, and other communication tools.

Commerce through Internet: It refers to buying and selling goods and services online.

Newsgroups (Usenet): It is an international discussion group that focuses on a particular topic and helps in gathering information about that topic.

Mailing Lists (List server): It refers to a large community of individuals who carry out active discussions, organized around topic-oriented forums that are distributed via e-mail and this method is known as mailing list.

3.1.3 Internetworking:

- Internetworking is the practice of connecting a computer network with other networks through the use of gateways that provide a common method of routing information packets between the networks. The resulting system of interconnected networks are called an internetwork, or simply an internet. Internetworking is a combination of the words inter ("between") and networking; not internet-working or international-network.

- The most notable example of internetworking is the Internet, a network of networks based on many underlying hardware technologies, but unified by an internetworking protocol standard, the Internet Protocol Suite, often also referred to as TCP/IP.

- The smallest amount of effort to create an internet (an internetwork, not the Internet), is to have two LANs of computers connected to each other via a router. Simply using either a switch or a hub to connect two local area networks together doesn't imply internetworking, it just expands the original LAN.

3.1.4 Internet Access

- Internet access is the process that enables individuals and organisations to connect to the Internet using computer terminals, computers, mobile devices, sometimes via computer networks. Once connected to the Internet, users can access Internet services, such as email and the World Wide Web. Internet service providers (ISPs) offer Internet access through various technologies that offer a wide range of data signaling rates (speeds).

- Consumer use of the Internet first became popular through dial-up Internet access in the 1990s. By the first decade of the 21st century, many consumers in developed nations used faster, broadband Internet access technologies. As of 2014, broadband was ubiquitous around the world, with a global average connection speed exceeding 4 Mbit/s.

Technologies for internet access:

When the Internet is accessed using a modem, digital data is converted to analog for transmission over analog networks such as the telephone and cable networks. A computer or other device accessing the Internet would either be connected directly to a modem that communicates with an Internet service provider (ISP) or the modem's Internet connection would be shared via a Local Area Network (LAN) which provides access in a limited area such as a home, school, computer laboratory, or office building.

a) Hardwired broadband access Technology

The term broadband includes a broad range of technologies, all of which provide higher data rate access to the Internet. The following are the hardwired broadband access techniques:

- Dial-up access
- Multilink dial-up,
- Integrated Services Digital Network
- Leased lines
- Cable Internet access
- Digital subscriber line (DSL, ADSL, SDSL, and VDSL)
- Fiber to the home
- Power-line Internet

b) Wireless broadband access Technology

Wireless broadband is used to provide both fixed and mobile Internet access with the following technologies:

- Satellite broadband
- Mobile broadband
- WiMAX
- Wireless ISP
- Local Multipoint Distribution Service

3.2 Using of Internet – understanding the basics, title, menu and tool bars, Address bar, Navigating web pages and Web sites and Printing.

3.2.1 Understanding the basics of Internet:

Internet

- The Internet is a network of computers spanning the globe. This communication structure is a system connecting people all around the world. A global Web of computers, the Internet allows individuals to communicate with each other. Often called the World Wide Web, the Internet provides a quick and easy exchange of information and is recognized as the central tool in this Information Age.

Internet Browser

- An Internet browser is a software program that enables you to access and navigate the Internet by viewing Web pages on your computer. The label Internet Browser describes

a software program that provides users with a graphical interface that allows them to connect to the Internet and “surf the Web.” Simply speaking, a browser is a software program that enables you to view Web pages on your computer.

- Internet Explorer and Firefox are commonly used for viewing the Internet. Internet Explorer and Firefox share many of the same functions, and it is possible to use both. There are other browsers available as well. It does not take many users long to develop a preference and “adopt” a browser. You may have already made the choice. Which are you using?

Web Site

- A site or area on the World Wide Web that is accessed by its own Internet address is called a Web site. A Web site can be a collection of related Web pages. Each Web site contains a home page and may also contain additional pages. Each Web site is owned and updated by an individual, company, or organization. Because the Web is a dynamically moving and changing entity, many Web sites change on a daily or even hourly basis.

Web Page

- A Web page can be explained as one area of the World Wide Web. Comparable to a page in a book, the basic unit of every Web site or document on the Web is a page. A Web page can be an article, an ordering page, or a single paragraph, and it is usually a combination of text and graphics.

Home Page

- The term home page has a couple of meanings. It is the Web page that your browser uses when it starts, and also the Web page that appears every time you open your browser. Clicking the home page icon on your browser screen will take you to the specific page you have set as your browser’s home page.

- Home page also refers to the main Web page out of a collection of Web pages. On each site, often you will see home page as a choice on a Menu Bar. Clicking on the word Home on a Web page will take you to the home or main page of that particular Web site

3.2.2 Title, menu and tool bars, address bar:

Title Bar

- The name of the Web site or title of the page you are viewing is found on the top left hand corner of your screen. Traditionally, this horizontal blue bar runs across the entire width of your screen. This blue bar that contains the name of the Web site is called the Title Bar. The Title Bar will serve as a trusty anchor, always letting your know where you are by sharing the title of the Web site you are visiting. This bar does not take you anywhere, but it always lets you know where you are.

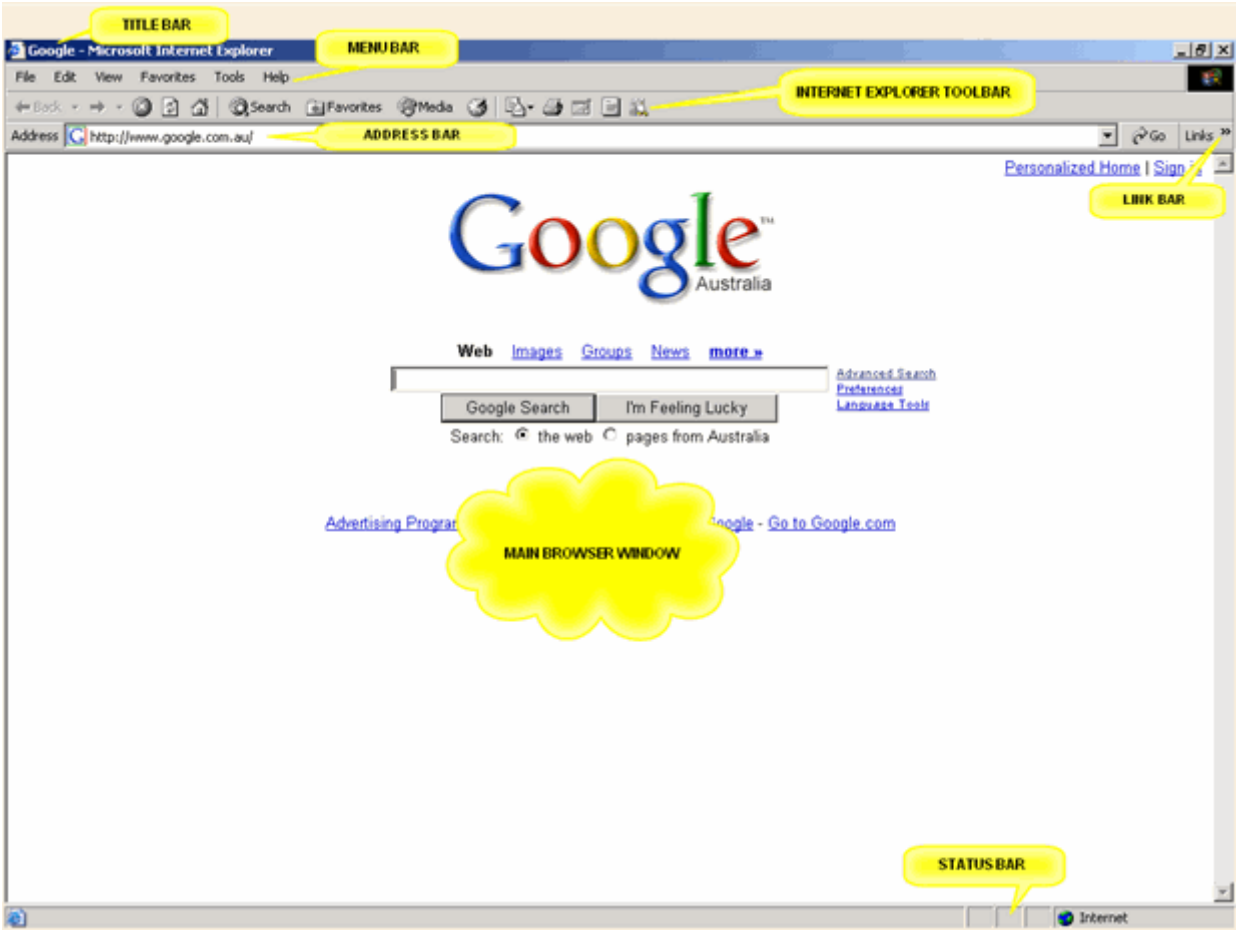
Menu Bar

- Underneath the Title Bar are other bars that can be used for moving around the Internet. If you are looking for quick and easy ways to navigate, the bars located at the top of your screen under the Title Bar will be helpful. One of the most useful bars is the Menu Bar. You will quickly appreciate each of the options found on the Menu Bar.

- The Menu Bar is the horizontal band that contains commands and options that can be chosen. In Internet Explorer, these selections are File, Edit, View, Favorites, Tools, and Help.

- Clicking on each of the items in the standard Menu Bar at the top of your page will drop down a menu that is a useful way to access the many features of the Internet Explorer program. The last menu item is the Help item. You will be surprised and relieved how often you will be able to click Help and find the answers you need.

- The Menu Bar is a very useful tool when trying to make your way around a Web site. Because the Menu Bar offers so many helpful functions, the quicker you master File, Edit, View, Favorites, Tools, and Help, the better. It does not take long to learn the purpose of each of these menu items that help you move around the Internet. The Main Window has many parts such as the Title Bar, Menu Bar, Toolbar, Address Bar, Link Bar, Main Browser Window as well as the Status Bar. These are shown in the diagram below



Address Bar

- The Address Bar is an excellent tool that can be used for navigating the Web. If you know the address of a page you want to visit, type the URL in the Address Bar. Then press Enter on the keyboard or click on the word Go on the right side of the Address Bar.

- The power of the computer really shines through with a feature called AutoComplete which is built into Internet Explorer. If you start typing a Web address that you have previously used, a list of matching addresses appears. The addresses of all the Web sites that you have visited are kept in the computer memory. Your browser will locate an address that you have previously typed in the Address Bar by searching for similar addresses trying to find a match. As you type each letter, the list is refined to match your typing. You can choose one of these addresses by simply moving your mouse over the name and clicking.

Tool Bar

The main toolbar is composed of eleven different buttons. Each of these buttons has a different function and purpose in Internet Explorer. The individual buttons will each be discussed in the following sections.

- ***The Back Button :*** This button will take you back to whatever document you were previously viewing. Pressing it immediately takes you back one document. If you have browsed many pages, or are well into a multi-page document, pressing it repeatedly will continue to back you up one page at a time. Once you reach your starting location, it will be greyed-out and unavailable.

- ***The Forward Button :*** This button will take you forward to the next document if you have previously browsed multiple documents and had then backed-up to the page you are currently viewing. (If you have not backed up at all, the forward button will be greyed-out) Pressing it repeatedly will continue to move you forward one page at a time. You can move forward until you reach the last page that you had browsed, at which time the forward button will be greyed-out.

- ***The Stop Button :*** The stop button stops ANY current operations by Internet Explorer. It will stop any type of file from loading. It can also be used to stop animations from continuing once a page is loaded. If you press it before a page has finished loading, the page will display everything it had finished loading before the stop button was pressed. If a document is completely loaded and there are no animations, movies, or other files still running, the stop button will have no immediate function.

- ***The Refresh Button :*** This button will reload the current document that you are viewing. It is useful if the page updates very frequently so that you can view these changes as soon as they are available. If you are loading a document and the transfer was interrupted, you can reload the full document again by clicking here.

- ***The Home Button :*** This button will return you to the page you have selected as the default start-up page for Internet Explorer. It will not take you back to the beginning of your web browsing, it will just return you to your home location from where you are. If you press back after reaching your home page, you will go back to the page you left after you hit the Home button.

- ***The Search Button :*** This button will take you to the page you have selected as the default Web search page for Internet Explorer. If you have not selected a page it will take you to Microsoft's default search page.

- ***The Favorites Button :*** This button will open up the Favorites menu. You can choose a favorite that you wish to go to from the list, add a favorite to the list, or organize your favorites from this menu.

- ***The Print Button :*** The print button will bring up a Print dialog box. In the box you can decide if you would like to print the contents of the page you are viewing, how many pages you will print, and also how many copies you will print. Keep in mind that if you try to print a page that is graphics intensive, you will need a printer that is capable of printing graphics. Also, the more graphics and pages a Web site has, the longer it will take to print.

- ***The Font Button :*** Pressing this button causes Internet Explorer to cycle through the available font sizes. This button is useful if the text is too small to read, or too large to fit comfortably in the window.

- ***The Mail Button :*** This button will open into a drop down menu from which you can select to read or send E-Mail. You can also open up your newsgroups from this menu.

- ***The Edit Button :*** This button will ONLY be on your toolbar if you have a Windows system Web editor (such as Microsoft Frontpage or Microsoft Word) installed on your computer. If you press this button, it will launch that editor and open the document you are currently viewing in it.

3.2.3 Navigating web pages and websites and printing:

Hyperlinks
An easy way to move around on the World Wide Web is by hyperlinks. Hyperlinks provide a connection between Web pages that allows for amazingly easy access to other Web pages. A link or hyperlink can be text, an icon, a picture, or an icon that moves a user from one Web page or Web site to another. A hyperlink has an unseen Web address imbedded in it.

Positioning your cursor on a hyperlink and clicking your mouse will take you to the Web page whose address is embedded in the hyperlink. You can tell that this text has a hyperlink hidden in it because it is a different color and because it is underlined. Just click on the hyperlinked word and presto – you go off to the world of the underlined word. Hyperlinks are a great way to easily find out more about a particular word or concept. There seems to be no end to the information on this Information Highway!

Underlined Link
A text link appears as an underlined word on the web page. When you click this underlined word, you will instantly jump from one place to another.

Navigating Within a Web Page
So far, our main focus has been moving from Web page to Web page or navigating between Web sites, but that is only half the picture. Once you have moved from Web site to Web site and selected a Web page you want to concentrate on, there are convenient ways to move around that particular page itself.

Often a Web page holds more information than can fit on one screen. A Web page appears aligned to the upper left hand corner of your screen. There is often information that you cannot see farther down after the last line on the screen. Sometimes there is also more information to the right of the screen.

Slider & Arrows
Scrolling is an easy way to navigate on a Web page. You can scroll up and down and side to side by using either the horizontal or vertical onscreen scroll bars on the bottom and right side of the screen. To scroll using the onscreen scroll bars, simply position your cursor on the slider on the scroll bar. Hold the mouse button down and drag the slider up and/or down on the vertical scroll bar (or side to side on the horizontal scroll bar). You can also position your cursor over the arrows at the top and the bottom of the vertical scroll bar (left and right sides of the horizontal scroll bar) to move one line at a time.

Using Arrow Keys
The keyboard holds some other choices for helping you move around a Web page. The first are the Page Up and Page Down keys on your keyboard. Pressing these keys while on a Web page, will move you up and/or down the screen one page at a time. The Arrow keys on the keyboard are convenient tools for moving the focus of your computer screen up, down, left, or right. These keys will move the screen more slowly, moving one line at a time.

Printing Pages
With all this navigating and exploring of Web pages, you have probably seen information that you would like to save in the old-fashioned way — with a printed-paper copy. As you viewed pages on the Web, did you find information that you'd like to save for future reference or share with others?

The computer world gives you the option of saving an entire Web page or any part of it: text, graphics, or links in printed form. Printing Web pages is very easy, thanks to that helpful Tool Bar. See the printer icon on the Tool Bar? That is the answer if you want a paper copy of a Web page you have found useful. A thoughtful gesture is to share Internet information by printing Web pages for people who don't have access to the Web or a computer.

To print a Web page, just click the printer icon on your Tool Bar. The page will print according to all your default options, which is usually what you want anyway.

There is another, more detailed way to print material from the Internet. Go to the Menu Bar and click on File. You will now see a dropdown menu offering a variety of choices, one of them being Print. Click Print. You will now be able to print a Web page, a portion of a Web page, or several copies of a Web page by making specific selections. You can select the printing options you want.

A nice way to double-check yourself is to preview how a Web page will look before you click the Print command; just click Print Preview.

3.3 Understanding the World Wide Web (WWW):

- The World Wide Web came into being in 1991, thanks to developer Tim Berners-Lee and others at the European Laboratory for Particle Physics, also known as Conseil European pour la Recherche Nucleure (CERN). The CERN team created the protocol based on hypertext that makes it possible to connect content on the Web with hyperlinks. Berners-Lee now directs the World Wide Web Consortium (W3C), a group of industry and university representatives that oversees the standards of Web technology.

- Early on, the Internet was limited to noncommercial uses because its backbone was provided largely by the National Science Foundation, the National Aeronautics and Space Administration, and the U.S. Department of Energy, and funding came from the government. But as independent networks began to spring up, users could access commercial Web sites without using the government-funded network. By the end of 1992, the first commercial online service provider, Delphi, offered full Internet access to its subscribers, and several other providers followed.

- The most widely used part of the Internet is the World Wide Web (often abbreviated "WWW" or called “the Web"). Its outstanding feature is hypertext. In most Web sites, certain words or phrases appear in text of a different color than the rest; often this text is also underlined. When you select one of these words or phrases, you will be transferred to the site or page that is relevant to this word or phrase. Sometimes there are buttons, images, or portions of images that are "clickable."

- The Internet is often confused with the World Wide Web. The misperception is that these two terms are synonymous. The Internet is the collection of the many different systems and protocols. The World Wide Web, developed in 1989, is actually one of those different protocols. As the name implies, it allows resources to be linked with great ease in an almost seamless fashion.

- The World Wide Web contains a vast collection of linked multimedia pages that is ever-changing. However, there are several basic components of the Web that allow users to communicate with each other. Below you will find selected components and their descriptions.

TCP/IP protocols

- In order for a computer to communicate on the Internet, a set of rules or protocols computers must follow to exchange messages was developed. The two most important protocols allowing computers to transmit data on the Internet are Transmission Control Protocol (TCP) and Internet Protocol (IP). With these protocols, virtually all computers can communicate with each other. For instance, if a user is running Windows on a PC, he or she can communicate with iPhones.

Domain name system

- An Internet address has four fields with numbers that are separated by periods or dots. This type of address is known as an IP address. Rather than have the user remember long strings of numbers, the Domain Name System (DNS) was developed to translate the numerical addresses into words. For example, the address fcit.usf.edu is really 131.247.120.10.

URLs

- Addresses for web sites are called URLs (Uniform Resource Locators). Most of them begin with http (HyperText Transfer Protocol), followed by a colon and two slashes. For example, the URL for the Florida Center for Instructional Technology is http://fcit.usf.edu/ .

- Some of the URL addresses include a directory path and a file name. Consequently, the addresses can become quite long. For example, the URL of a web page may be:
- http://fcit.usf.edu/holocaust/default.htm. In this example, "default.htm" is the name of the file which is in a directory named "holocaust" on the FCIT server at the University of South Florida.

Top-level domain

- Each part of a domain name contains certain information. The first field is the host name, identifying a single computer or organization. The last field is the top-level domain, describing the type of organization and occasionally country of origin associated with the address.

Top-level domain names include:

.com Commercial
.edu Educational
.gov US Government
.int Organization
.mil US Military
.net Networking Providers
.org Non-profit Organization

Domain name country codes include, but are not limited to:

.au Australia
.de Germany
.fr France
.nl Netherlands
.uk United Kingdom
.us United States

- Paying attention to the top level domain may give you a clue as to the accuracy of the information you find. For example, information on a "com" site can prove useful, but one should always be aware that the intent of the site may be to sell a particular product or service. Likewise, the quality of information you find on the "edu" domain may vary. Although many pages in that domain were created by the educational institutions themselves, some "edu" pages may be the private opinions of faculty and students. A common convention at many institutions is to indicate a faculty or student page with a ~ (tilde) in the address. For instance, http://fcit.usf.edu/~kemker/default.htm is a student's personal web page.

3.4 Searching Tools – World Search Engines, Search Directories and encyclopedias.

3.4.1 World Search Engines

- A search engine is a searchable database of Internet files which allows the user to enter keywords relating to particular topic and retrieve information about Internet sites containing those keywords.
- Search engines – webmasters and search engine optimization (SEO) professionals follow their guidelines for the highest possible rankings on them; paid search marketers pay to be featured on them; and users turn to them when they’re searching for answers, information, or entertainment.
- Search Engine Watch has been covering search engines since June 1997 and has watched the industry evolve to its current state. Over time, many search engines have come and gone, as users have spoken with their keyboards (and literally with their voices – thanks to voice search technology).
- In recent years, search market share has remained mostly unchanged – for much of the world, it’s Google followed by every other search engine (in the U.S. the “Big 5” search engines consist of Google, Bing, Yahoo, Ask.com and AOL, which combine for

hundreds of billions of searches every month). Meanwhile, many of the players have consolidated or have become footnotes in history.

These are the general world search engines:

Name	Language
Baidu	Chinese, Japanese
Bing	Multilingual
DuckDuckGo	Multilingual
Ecosia*	Multilingual
Exalead	Multilingual
Gigablast	English
Google	Multilingual
Munax	Multilingual
Qwant	Multilingual
Sogou	Chinese
Soso.com	Chinese
Yahoo!*	Multilingual
Yandex	Multilingual
Youdao	Chinese

1. **Google** – No need for further introductions. The search engine giant holds the first place in search with a stunning difference of 45% from second in place Bing. According to the latest **comscore report** (October 2012) 69.5% of searches were powered by Google and 25% by Bing. Google is also dominating the mobile/tablet search engine market share with 89%!

2. **Bing** – Bing is Microsoft's attempt to challenge Google in the area of search but despite their efforts they still did not manage to convince users that their search engine can produce better results than Google.

3. **Yahoo** – Since October 2011 Yahoo search is powered by Bing. Yahoo is still the most popular email provider and according to **reports** holds the third place in search.

4. **Ask.com** – Formerly known as Ask Jeeves, Ask.com receives approximately 3% of the search share. ASK is based on a question/answer format where most questions are answered by other users or are in the form of polls. It also has the general search functionality but the results returned lack quality compared to Google or even Bing and Yahoo.

5. **AOL**.com – According to **netmarketshare** the old time famous AOL is still in the top 10 search engines with a market share that is close to 0.6%. The AOL network includes many popular web sites like engadget.com, techchrunch.com and the huffingtonpost.com.

6. **Blekko**.com – **Blekko.com** was developed by ex-Googlers and they present themselves as the "spam free search engine". It is better suited for webmasters and SEO's who need more data for SEO purposes rather than normal users.

7. **Wolframalpha – wolframalpha** is different that all the other search engines. They market it as a Computational Knowledge Engine which can give you facts and data for a number of topics. It can do all sorts of calculations, for example if you enter *"mortgage 2000"* as input it will calculate your loan amount, interest paid etc. based on a number of assumptions.

8. **DuckDuckGo** – Has a number of advantages over the other search engines. It has a clean interface, it does not track users, it is not fully loaded with ads and has a number of very nice features (only one page of results, you can search directly other web sites etc). I am sure that some of the features of duckduckgo will be used by other search engines and with some proper funding duckduckgo can get a decent search engine market share.

9. **WayBackMachine** – **archive.org** is the internet archive search engine. You can use it to find out how a web site looked since 1996. It is very useful tool if you want to trace the history of a domain and examine how it has changed over the years.

10. **ChaCha.com** – According to **alexa** chacha.com is the 8th most popular search engine with a ranking position of 297 in the US. It is similar to ask.com where users can ask or answer a particular question. They also have a number of quizzes that can help you decide on a number of topics. It's not bad at all and the answers are precise and to the point. For example if you search "What is the best search engine?" you will get an answer that Google is the best and most popular search engine and Yahoo is on the second place.

These are the 10 best and most popular search engines on the Internet today. The list is by no means complete and for sure many more will be created in the future but as far as the first places are concerned, Google and Bing will hold the lead positions for years to come.

3.4.2 Search Directories and Encyclopedias

Search Directories perform the same function as search engines but they do not use computers to rank pages, they use people. People visit the submitted site and approve the site to a relevant directory. Yahoo is the best known search directory, although alot of people confuse it with a search engine. Here are some popular Search Directories:

- Yahoo - Most popular Directory
- DMOZ - Popular, public edited Directory
- 1st SPOT - General Directory
- The Net One - General Directory
- Ask Jeeves - Question related directory

Infomine
Browse and search scholarly Internet resources.
infomine.ucr.edu

Internet Scout Project
Locates high-quality resources of interest to researchers and educators.
scout.wisc.edu

ipl2
Searchable, subject-categorized directory of authoritative web sites.
ipl.org - Formerly: Internet Public Library & Librarians' Internet Index

Intute
Database of resources for research and education.
intute.ac.uk

Open Directory
Largest, most comprehensive human-edited directory of the web.
dmoz.org

RefSeek Reference Directory
Guide to the absolute best online reference resources.
refseek.com/directory

Virtual Reference Library
Information resources selected by the Toronto Public Library.
virtualreferencelibrary.ca

Yahoo! Directory
Pay-for-inclusion directory from Yahoo. Dates back to 1994.
dir.yahoo.com

3.4.3 ENCYCLOPEDIAS

An **encyclopedia** or **encyclopaedia** (also spelled **encyclopædia**, see spelling differences) is a type of reference work or compendium holding a comprehensive summary of information from either all branches of knowledge or a particular branch of knowledge. Encyclopedias are divided into articles or entries, which are usually accessed alphabetically by article name. Encyclopedia entries are longer and more detailed than those in most dictionaries. The following are the important online encyclopedias:

Answers.com
Encyclopedia aggregator and one-stop-shop for academic information.
www.answers.com - Aggregator

Britannica
Featuring 100,000 scholarly articles. Highly respected.
www.britannica.com - Encyclopedia

Catholic Encyclopedia
10,000 articles on Catholic history, interests, and doctrine.
www.newadvent.org/cathen - Religion

Columbia
Columbia's online encyclopedia offers 51,000 entries.
education.yahoo.com - Encyclopedia

Computer Desktop Encyclopedia
Featuring 20,000 topics and 2,500 images. Search field at bottom of page.
www.computerlanguage.com - Computers

Encyclopedia Mythica
Premier encyclopedia covering mythology, folklore, and religion.
www.pantheon.org - Mythology

Encyclopedia of Life
Database of scientific information on and photographs of known species.
www.eol.org - Biology

Encyclopedia of Philosophy
Hundreds of original contributions by philosophy experts and volunteers.
www.iep.utm.edu - Philosophy

Encyclopedia of Symbols
2,500 western signs and ideograms grouped by graphic characteristics.
www.symbols.com - Symbols

Encyclopedia Smithsonian
Online resources from the Smithsonian Institution.
www.si.edu - Encyclopedia

Europeana
Digital collection of scientific and cultural texts, images, videos, and sounds.
www.europeana.eu

Freebase
Open database of the world's information.
www.freebase.com - Encyclopedia

How Stuff Works
Source for easy-to-understand explanations of how things work.
www.howstuffworks.com - Science

Medline Medical Encyclopedia (by ADAM)
Diseases, symptoms, injuries, and more with photographs and illustrations.
www.nlm.nih.gov/medlineplus - Medical Encyclopedia

Reference.com
Encyclopedia aggregator. Sources include Columbia Encyclopedia and Wikipedia.
www.reference.com - Information Aggregator

Scholarpedia
Peer-reviewed open-access encyclopedia written by scholars from around the world.
www.scholarpedia.org

Stanford Encyclopedia of Philosophy
Hundreds of refereed encyclopedia entries.
plato.stanford.edu - Philosophy

Who2
Encyclopedia of more than 3,000 famous people and fictional characters.

www.who2.com - Famous People

Wikipedia
Find more than 2,000,000 entries, written and edited by users.
www.wikipedia.org - Encyclopedia

World Book
Traditional encyclopedia in print since 1917. Subscription required.
worldbook.com - Encyclopedia

World Digital Library
Archive of primary materials from cultures around the world.
wdl.org - Encyclopedia

3.5 Online safety – Spywares and viruses.

3.5.1 Spywares and Viruses

- Spyware is a type of virus that is specifically designed to steal information about your activity on your computer. Spyware writers have a number of different objectives, mainly fraudulent financial gain. Spyware can perform a number of illicit functions, from creating pop up advertisements to stealing your bank login details by taking screen shots of the sites you visit and even logging the keys you type. Spyware may also be self-replicating.

- Potentially, a virus could arrive on your computer in the form of a Trojan, it could replicate itself before moving on to another computer (a worm) and also be designed as a piece of spyware. Viruses and spyware are types of malware, which also includes rootkits, dishonest adware and scareware.

The Risks
Viruses and spyware can attack your computer via the following means:

- Opening infected email attachments such as .exe files.
- Opening infected files from web-based digital file delivery companies (for example Hightail - formerly called YouSendIt, and Dropbox).
- Visiting corrupt websites.
- Via the internet, undetected by the user (worms are an example of this).
- Macros in application documents (word processing, spreadsheets etc).
- USB connected devices (eg memory sticks, external hard drives, MP3 players).
- CDs/DVDs.

Viruses and spyware can cause very serious consequences including:

- Identity theft.
- Fraud.
- Deletion, theft and corruption of data.
- A slow or unusable computer.

Antivirus Software

- It is vital to keep your antivirus software up to date in order to provide the most complete protection. Thousands of new viruses are detected every year, to say nothing of the variants of new and existing ones. Each has a set of characteristics or 'signatures' that enable antivirus software manufacturers to detect them and produce suitable updates.

- Most antivirus software automatically downloads these updates (sometimes referred to as 'definitions') on a regular basis, as long as you are online and have paid your annual subscription (for a paid-for product). This should ensure protection against even the latest virus threats.

Antivirus software scans for viruses in a number of different ways:

- It scans incoming emails for attached viruses.
- It monitors files as they are opened or created to make sure they are not infected.
- It performs periodic scans of the files on your computer.

Some antivirus software also scans USB connected devices (eg memory sticks, external hard drives, MP3 players), as they are connecting. Some also highlights suspect websites. Antivirus software will not protect you against:

- Spam.
- Any kind of fraud or criminal activity online not initiated by a virus.
- A hacker trying to break into your computer over the internet.

Virus & Spyware Protection:

Apart from installing antivirus/antispyware software and keeping it updated, we recommend a number of other ways in which to keep your computer protected against viruses and spyware. After all, prevention is better than cure.

- Do not open any files attached to an email from an unknown, suspicious or untrustworthy source.
- Uninstall one antivirus program before you install another.
- Be careful with USB connected devices (eg memory sticks, external hard drives, MP3 players) as they are very common carriers of viruses.
- Be careful with CDs/DVDs as they can also contain viruses.
- Do not open any files from web-based digital file delivery companies (eg YouSendIt, Dropbox) that have been uploaded from an unknown, suspicious or untrustworthy source.
- Switch on macro protection in Microsoft Office applications like Word and Excel.
- Buy only reputable software from reputable companies.
- When downloading free software, do so with extreme caution.

*****

Unit-IV Bioinformatics

4.1 Introduction, scope and applications of bioinformatics.
4.2 Biological Databases – Protein and DNA sequences data bases; importance.
4.3 Genomics – Definitions, Pharmacogenomics, taxicogenomics, human genomics, prokaryotic and eukaryotic genomes and genome relationships.
4.4 Proteomics – Definitions, Transcriptomics and Metabolomics, Proteomics techniques (2D PAGE)
4.5 Computational Biology – Multiple Sequence Analysis and Phylogenetic alignment.

4.1 Introduction, scope and applications of bioinformatics.

4.1.1 Introduction

Bioinformatics was coined by Paulien Hogeweg and Ben Hesper in 1970. It was stated as "Study of Informatic processes in biotic systems". Basically bioinformatics deals with the information in the fields of Information Technology, Computer Science and Biology. Biologist performs research in laboratory and collects DNA and protein sequences, gene expressions etc. Computer Scientists are involved in developing algorithms, tools, softwares to store and analyze data. Bioinformaticians study biological questions by analyzing molecular data with various programs and tools. Today, bioinformatics is used in large number of fields such as microbial genome applications, biotechnology, waste cleanup, Gene Therapy etc. In this article an effort is made to provide brief information of applications of bioinformatics in the field of Medicine, Microbial Genome Application and Agriculture.

4.1.2 Scope of Bioinformatics

Bioinformatics uses advances in the area of computer science, information science, computer and information technology, communication technology to solve complex problems in life sciences and particularly in biotechnology. Data capture, data warehousing and data mining have become major issues for biotechnologists and biological scientists due to sudden growth in quantitative data in biology such as complete genomes of biological species including human genome, protein sequences, protein 3-D structures, metabolic pathways databases, cell line & hybridoma information, biodiversity related information.

Advancements in information technology, particularly the Internet, are being used to gather and access ever-increasing information in biology and biotechnology. Functional genomics, proteomics, discovery of new drugs and vaccines, molecular diagnostic kits and pharmacogenomics are some of the areas in which bioinformatics has become an integral part of Research & Development. The knowledge of multimedia databases, tools to carry out data analysis and modeling of molecules and biological systems on computer workstations as well as in a network environment has become essential for any student of Bioinformatics.

Bioinformatics, the multidisciplinary area, has grown so much that one divides it into molecular bioinformatics, organal bioinformatics and species bioinformatics. Issues related to biodiversity and environment, cloning of higher animals such as Dolly and Polly, tissue culture and cloning of plants have brought out that Bioinformatics is not only a support branch of science but is also a subject that directs future course of research in biotechnology and life sciences. The importance and usefulness of Bioinformatics is

realized in last few years by many industries. Therefore, large Bioinformatics R & D divisions are being established in many pharmaceutical companies, biotechnology companies and even in other conventional industry dealing with biological. Bioinformatics is thus rated as number one career in the field of biosciences.

4.1.3 Applications of Bioinformatics

In broad spectrum applications of bioinformatics is mainly used in the field of Medicine, Microbial Genome Applications and Agriculture.

1. Medicine

In the field of Medicine applications of bioinformatics is used for following areas:

a. Drug Discovery: The Idea of using X ray Crystallography in drug discovery emerged more than 30 years ago, when the first 3 dimensional structure of protein was determined. Within a decade, a radical change in drug design had begun, incarporating the knowledge of 3 dimensional structures of target protein into design process. Protein structure can influence drug discovery at every stage in design process. Classicaly, it is used in lead optimization, a process that uses structure to guide the chemical modification of a lead molecule to give an optimised fit in terms of shape, hydrogen bonds and other non - covalent interactions with the target.

b. Personal Medicine: Personalized medicine is a medical model that proposes the customization of healthcare, with all decisions and practices being tailored to the individual patient by use of genetic or other information. Practical application outside of long established considerations like a patient's family history, social circumstances, environment and behaviors are very limited so far and practically no progress has been made in the last decade. Personalized medicine research attempts to identify individual solutions based on the susceptibility profile of each individual. It is hoped that these fields will enable new approaches to diagnosis, drug development, and individualized therapy.

c. *Preventive Medicine**:* Preventive medicine or preventive care consists of measures taken to prevent diseases, (or injuries) rather than curing them or treating their symptoms. This contrasts in method with curative and palliative medicine, and in scope with public health methods (which work at the level of population health rather than individual health). Simple examples of preventive medicine include hand washing, breastfeeding, and immunizations.

d. *Gene Therapy:* Gene therapy is a novel form of drug delivery that enlists the synthetic machinery of the patient's cell to produce a therapeutic agent. It involves the efficient introduction of functional gene into the appropriate cells of the patient in order to produce sufficient amount of protein encoded by transferred gene (transgene) so as to precisely and permanently correct the disorder. Strategies of Gene Therapy are following:
- Gene addition
- Removal of harmful gene by antisense nucleotide or ribozymes
- Control of gene expression

2. Microbial Genome Applications

In the field of Microbial Genome Applications, applications of bioinformatics are used for following areas:

a. *Waste Cleanup*: In bioinformatics bacteria and microbes are identified which are

helpful in cleaning waste. *Deinococcus radiodurans* Bacterium is listed in the Guinness Book of World Records as "the world's toughest bacterium." This bacterium has the ability to repair damaged DNA and small fragments from chromosomes by isolating damage segments in a concentrated area [6]. This is because it has additional copies of its genome. Genes from other bacteria have been inserted into *D. radiodurans* for environmental cleanup. It was used to break down organic chemicals, solvents and heavy metals in radioactive waste sites .

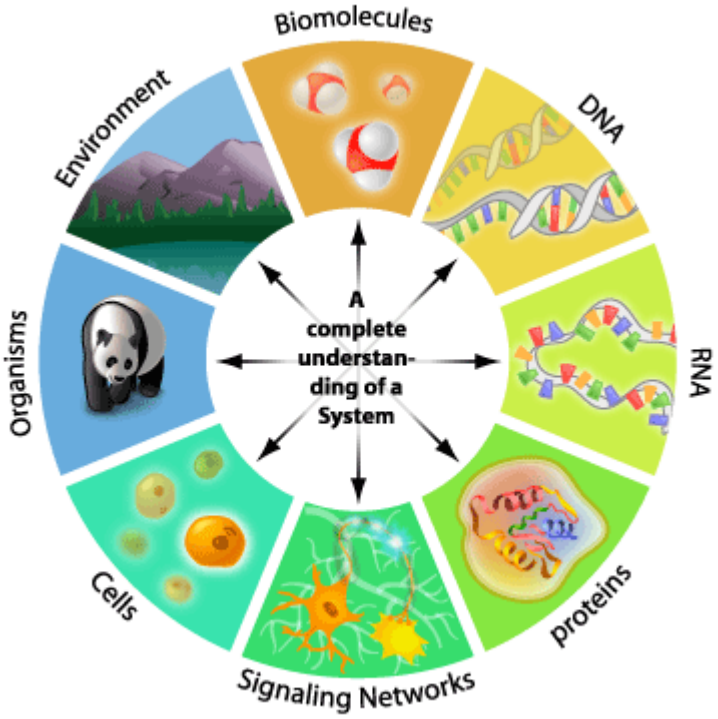


Fig: Bioinformatics in Life sciences

b. *Climate Change*: Climate change is caused by factors that include oceanic processes (such as oceanic circulation), variations in solar radiation received by Earth, plate tectonics and volcanic eruptions, and human-induced alterations of the natural world. By studying microorganisms genome scientists can begin to understand these microbes at a very fundamental level and isolated the genes that give them their unique abilities to survive under extreme conditions. *Rhodopseudomonas palustris* is a purple non-sulfur phototrophic bacterium commonly found in soils and water. It converts sunlight to cellular energy by absorbing atmospheric carbon dioxide and converting it to biomass. This microbe can also degrade and recycle a variety of aromatic compounds that comprise lignin, the main constituent of wood and the second most abundant polymer on earth [9]. *R. palustris* is acknowledged by microbiologists to be one of the most metabolically versatile bacteria ever described. Not only can it convert carbon dioxide gas into cell material but nitrogen gas into ammonia, and it can produce hydrogen gas. It grows both in the absence and presence of oxygen. In the absence of oxygen, it prefers to generate all its energy from light by photosynthesis

.

c. *Biotechnology*: The wide concept of "biotech" or "biotechnology" encompasses a wide range of procedures for modifying living organisms according to human purposes, going back to domestication of animals, cultivation of plants, and "improvements" to these through breeding programs that employ artificial selection and hybridization. Modern usage also includes genetic engineering as well as cell and tissue culture technologies. In the field of bioinformatics, biotechnology has identified organisms and microorganisms which can be very useful in dairy industry and food manufacturers. Lactococcus Lactis is

one of the most important micro-organisms involved in the dairy industry, it is a non-pathogenic rod-shaped bacterium that is critical for manufacturing dairy products like buttermilk, yogurt and cheese. This bacterium, is also used to prepare pickled vegetables, beer, wine, some breads and sausages and other fermented foods. Researchers anticipate that understanding the physiology and genetic make-up of this bacterium will prove invaluable for food manufacturers as well as the pharmaceutical industry, which is exploring the capacity of L. lactis to serve as a vehicle for delivering drugs.

d. *Alternative Energy*: Scientists are studying the genome of the microbe *Chlorobium tepidum*which has an unusual capacity for generating energy from light. *Chlorobium tepidum*is a thermophilic Gram-negative green sulfur bacterium isolated from a hot spring in New Zealand in which it forms a dense mat. The bacterium carries out photosynthesis in ways that are different from plants and other bacteria. Unlike plants, the green bacteria do not produce oxygen from photosynthesis. According to some researchers, photosynthesis may have its evolutionary origins in organisms like C. tepidum. Such species would have been able to harvest energy from light at a time when the Earth's atmosphere had little oxygen. In addition, the organisms' ability to grow in low-light environments may have helped them limit their exposure to UV irradiation, which was likely at higher levels in the early days of Earth.

3. Agriculture

In the field of Agriculture, applications of bioinformatics are in following areas:

a. *Crop Improvement*: Comparative genetics of the plant genomes has shown that the organisation of their genes has remained more conserved over evolutionary time than was previously believed. These findings suggest that information obtained from the model crop systems can be used to suggest improvements to other food crops. Arabidopsis thaliana (water cress) and *Oryza sativa* (rice) are examples of available complete plant genomes. Arabidopsis thaliana was the first plant to be sequenced and is considered the model species for investigating plant genetics and biology. There are many genes which are similar in all plants and the study of genes in a model organism like A. thaliana facillitates our understanding of gene expression and function in all plants. Furthermore, since animals and plants are both eukaryotes, many of the genes found in A. thaliana have homologs in animals. Arabidopsis has the smallest genome of any flowering plant, which is the main reason it was selected as a model organism for genome sequencing The DNA of Arabidopsis is made up of about 140 million bases, which are parcelled into five chromosomes. *Oryza sativa* (rice) is the most important crop for human consumption, providing staple food for more than half of the world population. *Oryza sativa* was the cereal selected to be sequenced as a priority and has gained the status "model organism". It has the smallest genome of all the cereals: 430 million nucleotides and it can serve as a model genome for one of the two main groups of flowering plants, the monocotyledons. Because it has been the subject of studies on yield, hybrid vigor, genetic resistance to disease and adaptive responses, scientists have taken advantage of the existence of a multitude of varieties that have adapted to a very wide range of environmental conditions, from dry soil in temperate regions to flooded cultures in tropical regions.

b. *Insect Resistance*: Genes from Bacillus thuringiensis that can control a number of serious pests have been successfully transferred to cotton, maize and potatoes. This new ability of the plants to resist insect attack means that the amount of insecticides being used can be reduced and hence the nutritional quality of the crops is increased. Bacillus thuringiensis is a pathogenic bacteria used for insect control. It is Gram-positive spore-forming, rod-shaped aerobic bacteria in the genus Bacillus. *B. thuringiensisis* an insecticidal bacterium, marketed worldwide for control of many important plant pests -

mainly caterpillars of the Lepidoptera (butterflies and moths) but also mosquito larvae, and simuliid blackflies that vector river blindness in Africa. B. thuringiensis products represent about 1% of the total 'agrochemical' market (fungicides, herbicides and insecticides) across the world. The commercial B. thuringiensis products are powders containing a mixture of dried spores and toxin crystals. They are applied to leaves or other environments where the insect larvae feed. The toxin genes have also been genetically engineered into several crop plants.

c. *Improve Nutritional Quality*: Scientists have recently succeeded in transferring genes into rice to increase levels of Vitamin A, iron and other micronutrients. This work could have a profound impact in reducing occurrences of blindness and anaemia caused by deficiencies in Vitamin A and iron respectively. Scientists have inserted a gene from yeast into the tomato, and the result is a plant whose fruit stays longer on the vine and has an extended shelf life. One little gene may be all that stands between a fresh, juicy, homegrown tomato and its bland, store-bought counterpart. Biologists announced that they've identified the gene that controls the ripening process in the humble fruit. If this "rin" gene can be manipulated effectively, scientists will be able to create breeds of tomatoes that will be more flavorful even after the long journey from the vine to the produce department. Today, tomatoes are plucked from the vine early, when still green and firm, to ensure that they survive shipping without bruising and rotting. Picking tomatoes early means they have less chance to develop flavor, color, and nutrients naturally. By manipulating the "rin" gene, scientists will be able to slow the ripening process, letting the tomato develop on the vine for longer - but still keeping it firm enough to ship safely. The scientists responsible for the "rin" gene findings are from the U.S. Department of Agriculture and the Boyce Thompson Institute for Plant Research, on the campus of Cornell University. They hope that their technique may also be applied to other fruits - such as strawberries, bananas, bell peppers, and melons - which suffer from the same shipping and storage complications.

4.2 Biological Databases – Protein and DNA sequences data bases; importance.

4.2.1 Introduction to Biological Databases:

- **Biological databases** are libraries of life sciences information, collected from scientific experiments, published literature, high-throughput experiment technology, and computational analysis. They contain information from research areas including genomics, proteomics, metabolomics, microarray gene expression, and phylogenetics. Information contained in biological databases includes gene function, structure, localization (both cellular and chromosomal), clinical effects of mutations as well as similarities of biological sequences and structures.
- Databases are broadly classified as ***generalized databases*** and **specialized databases.** Structural organisation of DNA, protein, carbohydrates are included under *generalized databases*. Databases of Expressed Sequence Tags (ESTs), Genome Survey Sequences (GSS), Single Nucleotide Polymorphisms (SNPs) sequence Tagged sites (STSs). RNA databases are included under *specialized data bases*.

  Generalized databases contain sequence database and structure databases.

  a. *Sequence databases* are the sequence records of either nucleotides or amino acids. The former is the nucleic acid databases and the latter are the protein sequence databases.

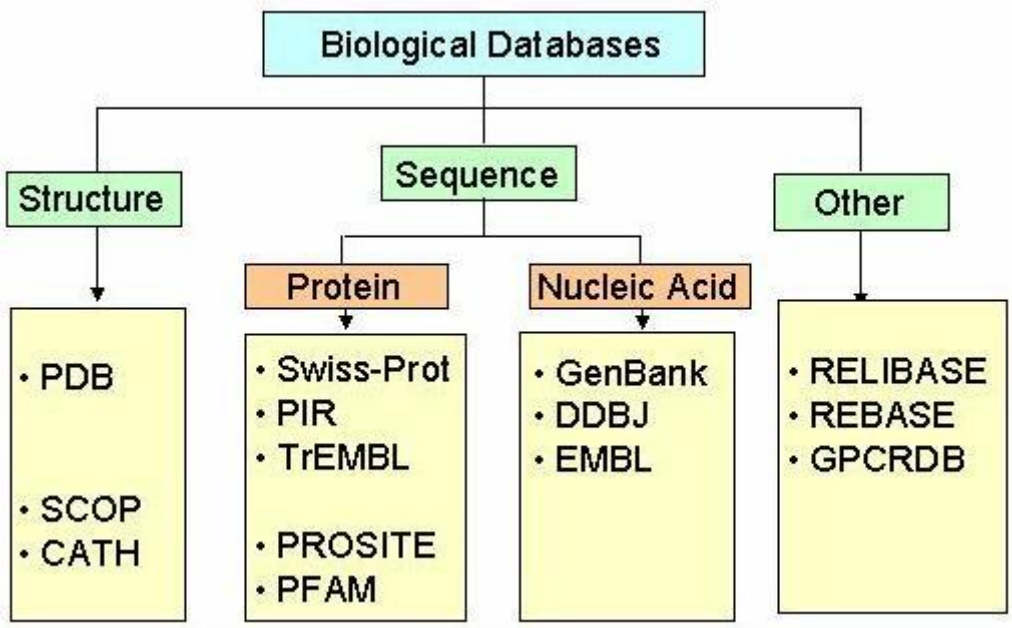
b. *Structure databases* are the individual records of macromolecular structures. The nucleic acid databases are again classified into primary databases and secondary databases.

- Primary databases contain the data in their original form taken as such from the source eg., Genebank (NCBI/USA) Protein, SWISS-PROT (Switzerland), Protein 3D structure etc.

DNA (nucleotide)		Protein	
EMBL	UK	PIR	US
GenBank	US	MIPS	Germany
DDBJ	Japan	Swiss-Prot	Swiss
Celera	Celera	TrEMBL	Swiss
		NRL_3D	US
		GenPept	US

1: List of primary sequence databases and their locations.

- Secondary databases also called as value added databases contain annotated data and information eg., OMIN Online Mendelian Inheritance in Man. GDB - Genome Database Human.



4.2.2 Protein sequence data base

The protein sequence databases elucidate the high level annotations such as the description of the protein functions; their domain structure (configuration), amino acid sequence, post-translational modifications, variants etc. SWISS-PROT groups at SIB (Swiss Institute of Bioinformatics) and EBI (European Bioinformatics Institute) have developed the protein sequence databases. SWISS-PROT is revealed at http://www.expasy.ch/sprot-top.html.

To exploit the various resources fully, it is essential to distinguish between them and to identify the types of data they contain. Universal protein databases cover proteins from all species whereas specialized data collections contain information about a particular protein family or group of proteins, or related to a specific organism.

Universal protein sequence databases can be further subdivided into two categories: ***sequence repositories***, in which data are stored with little or no manual intervention in the creation of the records; and ***expertly curated databases***, in which the original data

are enhanced by the addition of further information. In the following, we present the current status of the leading protein sequence databases.

a) Sequence repositories

- Several protein sequence databases act as repositories of protein sequences. These databases add little or no additional information to the sequence records they contain and generally make no effort to provide a non-redundant collection of sequences to users. The following are the sequence repositories.

i) NCBI's Entrez Protein:

NCBI's Entrez Protein (http://www.ncbi.nlm.nih.gov/ entrez/query.fcgi?db-Protein) is another example of a sequence repository. The database contains sequence data translated from the nucleotide sequences of the DDBJ/EMBL/GenBank database as well as sequences from Swiss-Prot

ii) RefSeq:

A more ambitious approach is taken by the Reference Sequence (RefSeq) collection produced by the NCBI (http://www.ncbi.nih.gov/RefSeq). The aim of the project is to provide a non-redundant collection of reference protein sequences. RefSeq sequences exist for a limited set of species, including approximately 1100 viruses and 150 bacteria, as well as a small number of higher organisms, such as human, mouse, rat, zebra fish, honeybee, sea urchin, cow, and several important plant species.

b) Curated databases

- The curated databases enrich the sequence data by adding additional information, which gets validated by expert biologists before being added to the databases to ensure that the data in these collections can be considered to be highly reliable. The following are the Curated databases:

i) PIR-PSD

The oldest universal curated protein sequence database is the Protein Information Resource Protein Sequence Database (PIR-PSD) (http://pir.georgetown.edu/). It compiles comprehensive, non-redundant protein sequence data, organized by superfamily and family, and annotated with functional, structural, bibliographic and genetic data.

ii) Swiss-Prot

The leading universal curated protein sequence database is Swiss-Prot (http://www.ebi.ac.uk/swissprot/index.html), which contained as of November 2003 (release 42.6) 140 000 curated sequence entries from over 8300 different species. The database is non-redundant, which means that all reports for a given protein are merged into a single entry, and is highly integrated with other databases.

iii) TrEMBL

To produce a fully curated Swiss-Prot entry is a highly labor-intensive process and is the rate-limiting step in the growth of the database because more new sequences are submitted than can be efficiently annotated manually and integrated into the database. To address this, the TrEMBL (Translation from EMBL) database (http://www.ebi.ac.uk/trembl/) was introduced to make new sequences available as quickly as possible.

iv) UniProt: the next generation of protein sequence databases

One of the most significant developments with regard to protein sequence databases is the recent decision by the National Institutes of Health to award a

grant to combine the Swiss-Prot, TrEMBL and PIR-PSD databases into a single resource, UniProt (http://www.uniprot.org).

v) The UniProt knowledgebase (UNIPROT)

Swiss-Prot, TrEMBL and PIR-PSD have been merged to form the UniProt knowledgebase. All suitable PIR-PSD sequences that are missing from Swiss-Prot + TrEMBL were incorporated into UniProt. Bi-directional cross- references between Swiss-Prot + TrEMBL and PIR-PSD were created to allow the easy tracking of the PIR-PSD entries. The transfer into UniProt of references and experimentally verified data present in PIR but missing from Swiss-Prot + TrEMBL is ongoing.

4.2.3 DNA or Nucleic Acid sequence data bases;

- The databases EMBL, GenBank, and DDBJ are the **three primary nucleotide sequence databases**: They include sequences submitted directly by scientists and genome sequencing group, and sequences taken from literature and patents. There is comparatively little error checking and there is a fair amount of redundancy.
- The entries in the EMBL, GenBank and DDBJ databases are **synchronized** on a daily basis, and the accession numbers are managed in a consistent manner between these three centers.
- The nucleotide databases have reached such large sizes that they are available in **subdivisions** that allow searches or downloads that are more limited, and hence less time-consuming. For example, GenBank has currently 17 divisions.
- There are **no legal restrictions** on the use of the data in these databases. However, there are patented sequences in the databases. The following are the DNA sequence databases.

- **EMBL. www.ebi.ac.uk/embl/**

  The EMBL (European Molecular Biology Laboratory) nucleotide sequence database is maintained by the European Bioinformatics Institute (EBI) in Hinxton, Cambridge, UK. As of 16 Jan 2001, it contained 10,378,022 records with a total of 11,302,156,937 bases; see the EMBL DB statistics page. It can be accessed and searched through the SRS system at EBI, or one can download the entire database as flat files. An example of what an entry looks like is given for the human raf oncogene protein, ID: HSRAFR.

- **GenBank. www.ncbi.nlm.nih.gov/Genbank/**

  The GenBank nucleotide database is maintained by the National Center for Biotechnology Information (NCBI), which is part of the National Institute of Health (NIH), a federal agency of the US government. It can be accessed and searched through the Entrez system at NCBI, or one can download the entire database as flat files. An example of what an entry looks like is given for the human raf oncogene protein, Locus: HSRAFR.

- **DDBJ www.ddbj.nig.ac.jp**

  The DNA Data Bank of Japan began as a collaboration with EMBL and GenBank. It is run by the National Institute of Genetics. One can search for entries by accession number, FASTA/BLAST, keywords and regular expressions.

- ***Other DNA sequence databases***

  The following databases contain subsets of the EMBL/GenBank databases. Some also contain more information or links than the primary ones, or have a different organization of the data to better suit some specific purpose. However, the nucleotide sequences themselves should always be available in the EMBL/GenBank databases. In this sense, the databases below are secondary databases.

- **UniGene. www.ncbi.nlm.nih.gov/UniGene/**
  The UniGene system attempts to process the GenBank sequence data into a non-redundant set of gene-oriented clusters. Each UniGene cluster contains sequences that represent a unique gene, as well as related information such as the tissue types in which the gene has been expressed and map location.

- **SGD genome-www.stanford.edu/Saccharomyces/**
  The Saccharomyces Genome Database (SGD) is a scientific database of the molecular biology and genetics of the yeast Saccharomyces cerevisiae.

- **EBI Genomes. www.ebi.ac.uk/genomes/**
  This web site provides access and statistics for the completed genomes, and information about ongoing projects.

- **Genome Biology. www.ncbi.nlm.nih.gov/Genomes/**
  The Genome Biology site at NCBI contains information about the available complete genomes.

- **Ensembl. www.ensembl.org**
  Ensembl is a joint project between EMBL-EBI and the Sanger Centre to develop a software system which produces and maintains automatic annotation of eukaryotic genomes.

4.2.4 Importance of Biological Databases

Databases are the system which is used to store, search and retrieve any type of data. Biological databases are libraries of life sciences information, collected from scientific experiments, published literature, high throughput experiment technology, and computational analyses of various research areas like genomics, proteomics, metabolomics, microarray gene expression, phylogenetic information, clinical trials. Relational database concepts of computer science and information retrieval concepts of digital libraries are important for understanding biological databases.

Biological database design, development, and long-term management are core areas of the discipline of bioinformatics (http://en.wikipedia.org/wiki/Biological_database). Due to increase in availability of genome sequences of various organisms in recent years, the biological data has accumulated with exponential rate. This created a necessary demand to manage, store and efficiently retrieve this biological wealth (Rene Witte and Christopher J.O. Baker, 2005).

The recent manipulation of large number of biological databases and improved bioinformatics tools has improved the better understanding of the biological systems.

The purposes of these biological databases are:-

- Make information available globally
- Systematic results from experiments and analysis
- Non-redundancy and redundancy deduction
- Accuracy
- Reference to literature
- Bioinformatics issues
- Database design and implementations
- Consistency

- Cross- references
- Tools for analyzing, querying and visualization
- Data mining

Generally, Biological databases are classified as Primary, Secondary and Composite databases. Primary database has experimental results in database. Secondary database has the results of analysis of primary database. Composite database combined with various primary database sources and reduce the risk of searching multiple resources. The two main functions of the biological databases are:

(a) To make available of medical databases to the research community and
(b) To make available of medical databases in a computer readable form.

4.3 Genomics – Definitions, Pharmacogenomics, taxicogenomics, human genomics, prokaryotic and eukaryotic genomes and genome relationships.

4.3.1 Genomics-Introduction:

- **Genomics** is a discipline in genetics that applies recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble, and analyze the function and structure of genomes (the *complete* set of DNA within a single cell of an organism).[1][2] Advances in genomics have triggered a revolution in discovery-based research to understand even the most complex biological systems such as the brain.[3] The field includes efforts to determine the entire DNA sequence of organisms and fine-scale genetic mapping. The field also includes studies of intragenomic phenomena such as heterosis, epistasis, pleiotropy and other interactions between loci and alleles within the genome.[4] In contrast, the investigation of the roles and functions of single genes is a primary focus of molecular biology or genetics and is a common topic of modern medical and biological research. Research of single genes does not fall into the definition of genomics unless the aim of this genetic, pathway, and functional information analysis is to elucidate its effect on, place in, and response to the entire genomes networks.

4.3.2 Pharmacogenomics (PGx)

- Pharmacogenomics is the study of how genes affect a person's response to drugs. This relatively new field combines pharmacology (the science of drugs) and genomics (the study of genes and their functions) to develop effective, safe medications and doses that will be tailored to a person's genetic makeup.
- Many drugs that are currently available are "one size fits all," but they don't work the same way for everyone. It can be difficult to predict who will benefit from a medication, who will not respond at all, and who will experience negative side effects (called adverse drug reactions). Adverse drug reactions are a significant cause of hospitalizations and deaths in the United States. With the knowledge gained from the Human Genome Project, researchers are learning how inherited differences in genes affect the body's response to medications. These genetic differences will be used to predict whether a medication will be effective for a particular person and to help prevent adverse drug reactions.

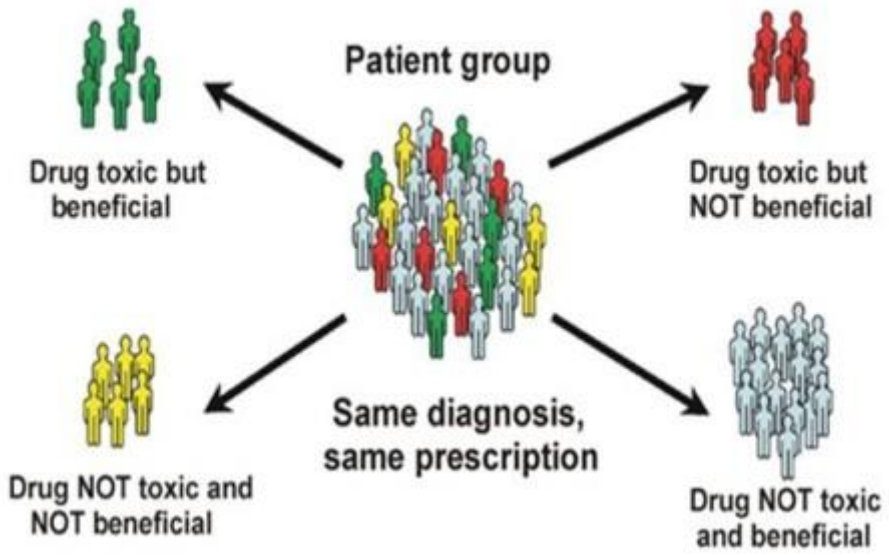


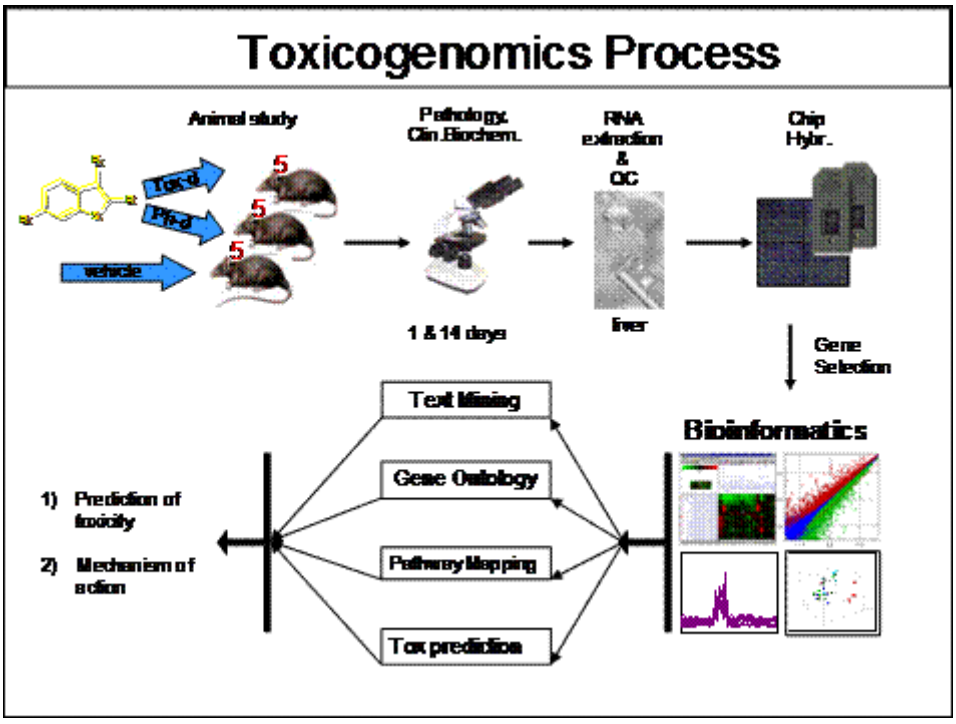
Figure. PGx (Pharmacogenomics) Testing

- PGx testing data indicates how your body processes medications based on your genes. If your on a medication that your body image001doesn't metabolize normally, you should seek medical advice. Recent industry studies have shown that the clinical effectiveness provided by PGx testing leads to reduced adverse drug events, improved clinical results, and lower healthcare costs. So, whether your starting a new medication or looking to improve your current medication therapy, PGx test reports can help you and your healthcare prescribers determine the right drug, at the right dose.
- The field of pharmacogenomics is still in its infancy. Its use is currently quite limited, but new approaches are under study in clinical trials. In the future, pharmacogenomics will allow the development of tailored drugs to treat a wide range of health problems, including cardiovascular disease, Alzheimer disease, cancer, HIV/AIDS, and asthma.

4.3.3 Taxicogenomics

- **Toxicogenomics** is a field of science that deals with the collection, interpretation, and storage of information about gene and protein activity within particular cell or tissue of an organism in response to toxic substances. Toxicogenomics combines toxicology with genomics or other high throughput molecular profiling technologies such as transcriptomics, proteomics and metabolomics.[1][2] Toxicogenomics endeavors to elucidate molecular mechanisms evolved in the expression of toxicity, and to derive molecular expression patterns (i.e., molecular biomarkers) that predict toxicity or the genetic susceptibility to it.
- In pharmaceutical research toxicogenomics is defined as the study of the structure and function of the genome as it responds to adverse xenobiotic exposure. It is the toxicological subdiscipline of pharmacogenomics, which is broadly defined as the study of inter-individual variations in whole-genome or candidate gene single-nucleotide polymorphism maps, haplotype markers, and alterations in gene expression that might correlate with drug responses (Lesko and Woodcock 2004, Lesko et al. 2003). Though the term toxicogenomics first appeared in the literature in 1999 (Nuwaysir et al.) it was already in common use within the pharmaceutical industry as its origin was driven by marketing strategies from vendor companies. The term is still not universal accepted, and others have offered alternative terms such as chemogenomics to describe essentially the same area.

- A typical toxicogenomics experiment is shown in the picture below. Groups of 5 rats are treated with a vehicle (placebo), a low pharmaceutical dose of a compound or a high toxic dose. After 1 and 14 days rats are sacrified and classical clinical chemistry parameters are measured and histopathology assessed. Parts of the livers are isolated and total RNA purified for subsequent chip hybridizations. Raw data is stored in a database and bioinformatics tools are used to detect differentially expressed genes and to characterize the gene set. Together with clinical chemistry and histopathology data toxicogenomics results will help to assess possible side effects of compounds and understand the mechanism of toxicity.



4.3.4 Human genome:

- The **human genome** is the complete set of nucleic acid sequence for humans (*Homo sapiens*), encoded as DNA within the 23 chromosome pairs in cell nuclei and in a small DNA molecule found within individual mitochondria. Human genomes include both protein-coding DNA genes and noncoding DNA. Haploid human genomes, which are contained in germ cells (the egg and sperm gamete cells created in the meiosis phase of sexual reproduction before fertilization creates a zygote) consist of three billion DNA base pairs, while diploid genomes (found in somatic cells) have twice the DNA content. While there are significant differences among the genomes of human individuals (on the order of 0.1%),[1] these are considerably smaller than the differences between humans and their closest living relatives, the chimpanzees (approximately 4%[2]) and bonobos.
- The Human Genome Project produced the first complete sequences of individual human genomes, with the first draft sequence and initial analysis being published on February 12, 2001.[3] The human genome was the first of all vertebrates to be completely sequenced. As of 2012, thousands of human genomes have been completely sequenced, and many more have been mapped at lower levels of resolution. The resulting data are used worldwide in biomedical science, anthropology, forensics and other branches of science. There is a widely held expectation that genomic studies will lead to advances in the diagnosis and treatment of diseases, and to new insights in many fields of biology, including human evolution.

- Although the sequence of the human genome has been (almost) completely determined by DNA sequencing, it is not yet fully understood. Most (though probably not all) genes have been identified by a combination of high throughput experimental and bioinformatics approaches, yet much work still needs to be done to further elucidate the biological functions of their protein and RNA products. Recent results suggest that most of the vast quantities of noncoding DNA within the genome have associated biochemical activities, including regulation of gene expression, organization of chromosome architecture, and signals controlling epigenetic inheritance.

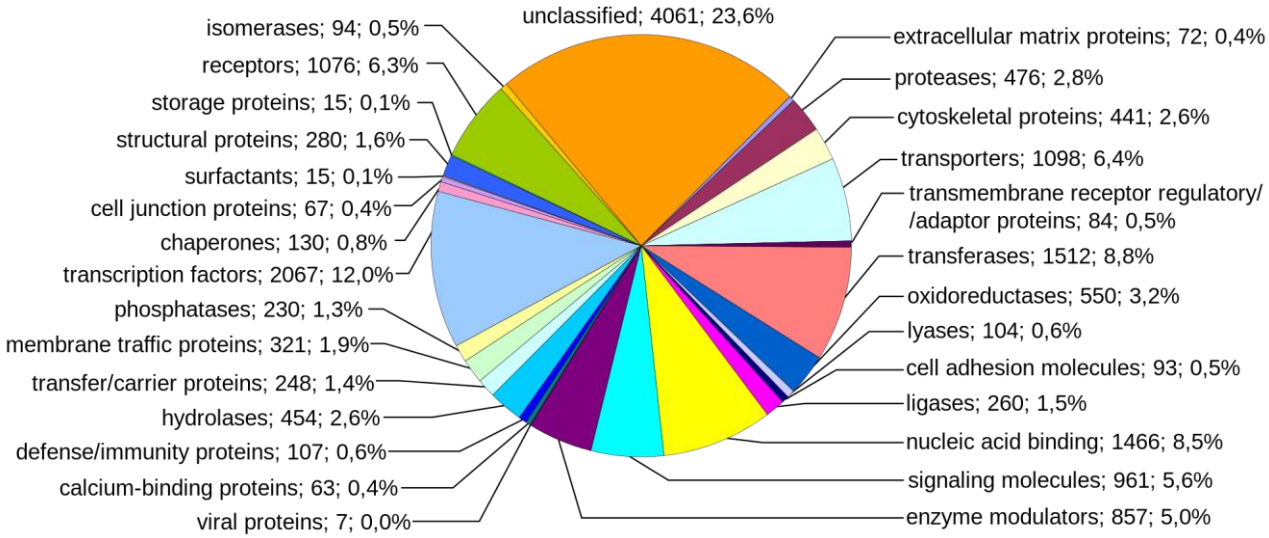


Figure. Human genes categorized by function of the transcribed proteins, given both as number of encoding genes and percentage of all genes

- There are an estimated 20,000-25,000 human protein-coding genes. The estimate of the number of human genes has been repeatedly revised down from initial predictions of 100,000 or more as genome sequence quality and gene finding methods have improved, and could continue to drop further.[4][5] Protein-coding sequences account for only a very small fraction of the genome (approximately 1.5%), and the rest is associated with non-coding RNA molecules, regulatory DNA sequences, LINEs, SINEs, introns, and sequences for which as yet no function has been determined.

4.3.5 Prokaryotic and Eukaryotic genomes

Biologists recognize that the living world comprises two types of organism:

1. Eukaryotes, whose cells contain membrane-bound compartments, including a nucleus and organelles such as mitochondria and, in the case of plant cells, chloroplasts. Eukaryotes include animals, plants, fungi and protozoa.
2. Prokaryotes, whose cells lack extensive internal compartments. There are two very different groups of prokaryotes, distinguished from one another by characteristic genetic and biochemical features:
   a. the bacteria, which include most of the commonly encountered prokaryotes such as the gram-negatives (e.g. *E.* coli), the gram-positives (e.g. *Bacillus subtilis*), the cyanobacteria (e.g. *Anabaena*) and many more;
   b. the archaea, which are less well-studied, and have mostly been found in extreme environments such as hot springs, brine pools and anaerobic lake bottoms.

A) Genomes of eukaryotes

- Humans are fairly typical eukaryotes and the human genome is in many respects a good model for eukaryotic genomes in general. All of the eukaryotic nuclear genomes that have been studied are, like the human version, divided into two or more linear DNA molecules, each contained in a different chromosome; all eukaryotes also possess smaller, usually circular, mitochondrial genomes. The only general eukaryotic feature not illustrated by the human genome is the presence in plants and other photosynthetic organisms of a third genome, located in the chloroplasts.
- Although the basic physical structures of all eukaryotic nuclear genomes are similar, one important feature is very different in different organisms. This is genome size, the smallest eukaryotic genomes being less than 10 Mb in length, and the largest over 100 000 Mb. As can be seen in Table 2.2, this size range coincides to a certain extent with the complexity of the organism, the simplest eukaryotes such as fungi having the smallest genomes, and higher eukaryotes such as vertebrates and flowering plants having the largest ones. This might appear to make sense as one would expect the complexity of an organism to be related to the number of genes in its genome - higher eukaryotes need larger genomes to accommodate the extra genes. However, the correlation is far from precise: if it was, then the nuclear genome of the yeast *S. cerevisiae*, which at 12 Mb is 0.004 times the size of the human nuclear genome, would be expected to contain 0.004 × 35 000 genes, which is just 140. In fact the *S. cerevisiae* genome contains about 5800 genes.

Species	Genome size (Mb)
Fungi	
Saccharomyces cerevisiae	12.1
Aspergillus nidulans	25.4
Protozoa	
Tetrahymena pyriformis	190
Invertebrates	
Caenorhabditis elegans	97
Drosophila melanogaster	180
Bombyx mori (silkworm)	490
Strongylocentrotus purpuratus (sea urchin)	845
Locusta migratoria (locust)	5000
Vertebrates	
Takifugu rubripes (pufferfish)	400
Homo sapiens	3200
Mus musculus (mouse)	3300
Plants	
Arabidopsis thaliana (vetch)	125
Oryza sativa (rice)	430
Zea mays (maize)	2500
Pisum sativum (pea)	4800
Triticum aestivum (wheat)	16 000
Fritillaria assyriaca (fritillary)	120 000

B) Genomes of prokaryotes

- Prokaryotic genomes are very different from eukaryotic ones. There is some overlap in size between the largest prokaryotic and smallest eukaryotic genomes, but on the whole prokaryotic genomes are much smaller. For example, the *E. coli* K12 genome is just 4639 kb, two-fifths the size of the yeast genome, and has only 4405 genes.
- The physical organization of the genome is also different in eukaryotes and prokaryotes. The traditional view has been that an entire prokaryotic genome is contained in a single circular DNA molecule. As well as this single 'chromosome', prokaryotes may also have additional genes on independent smaller, circular or linear DNA molecules called plasmids.
- Genes carried by plasmids are useful, coding for properties such as antibiotic resistance or the ability to utilize complex compounds such as toluene as a carbon source, but plasmids appear to be dispensable - a prokaryote can exist quite effectively without them. We now know that this traditional view of the prokaryotic genome has been biased by the extensive research on *E. coli*, which has been accompanied by the mistaken assumption that *E. coli* is a typical prokaryote.
- In fact, prokaryotes display a considerable diversity in genome organization, some having a unipartite genome, like *E. coli,* but others being more complex. *Borrelia burgdorferi* B31, for example, has a linear chromosome of 911 kb, carrying 853 genes, accompanied by 17 or 18 linear and circular molecules, which together contribute another 533 kb and at least 430 genes. Multipartite genomes are now known in many other bacteria and archaea.

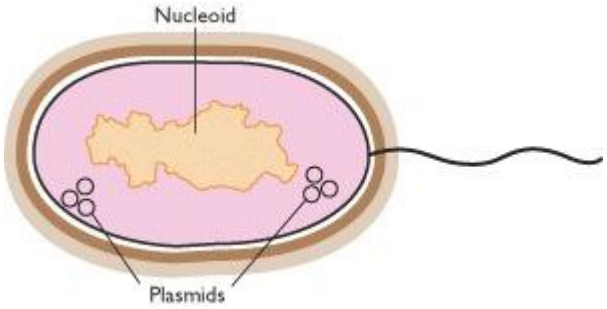


Figure. Plasmids are small circular DNA molecules that are found inside some prokaryotic cells

- In one respect, *E. coli* is fairly typical of other prokaryotes. After our discussion of eukaryotic gene organization, it will probably come as no surprise to learn that prokaryotic genomes are even more compact than those of yeast and other lower eukaryotes.
- A 50-kb segment of the *E. coli* K12 genome immediately obvious that there are more genes and less space between them, with 43 genes taking up 85.9% of the segment. Some genes have virtually no space between them: *thrA* and *thrB*, for example, are separated by a single nucleotide, and *thrC* begins at the nucleotide immediately following the last nucleotide of *thrB*. These three genes are an example of an operon, a group of genes involved in a single biochemical pathway (in this case, synthesis of the amino acid threonine) and expressed in conjunction with one another.
- Operons have been used as model systems for understanding how gene expression is regulated. In general, prokaryotic genes are shorter than their eukaryotic counterparts, the average length of a bacterial gene being about two-thirds that of a eukaryotic gene, even after the introns have been removed from the latter. Bacterial genes appear to be slightly longer than archaeal ones.

4.4 Proteomics – Definitions, Transcriptomics and Metabolomics, Proteomics techniques (2D PAGE)

4.4.1 Proteomics Introduction:

- **Proteomics** is the large-scale study of proteins, particularly their structures and functions. Proteins are vital parts of living organisms, as they are the main components of the physiological metabolic pathways of cells. The term *proteomics* was first coined in 1997 to make an analogy with genomics, the study of the genome. The word *proteome* is a portmanteau of *prote*in and gen*ome*, and was coined by Marc Wilkins in 1994 while working on the concept as a PhD student.
- The ***proteome*** is the entire set of proteins, produced or modified by an organism or system. This varies with time and distinct requirements, or stresses, that a cell or organism undergoes. Proteomics is an interdisciplinary domain that has benefited greatly from the genetic information of the Human Genome Project; it is also emerging scientific research and exploration of proteomes from the overall level of intracellular protein composition, structure, and its own unique activity patterns. It is an important component of functional genomics.
- Using the more inclusive definition of proteomics, many different areas of study are now grouped under the rubric of proteomics (Fig.). These include protein-protein interaction studies, protein modifications, protein function, and protein localization studies to name a few. The aim of proteomics is not only to identify all the proteins in a cell but also to create a complete three-dimensional (3-D) map of the cell indicating where proteins are located. These ambitious goals will certainly require the involvement of a large number of different disciplines such as molecular biology, biochemistry, and bioinformatics. It is likely that in bioinformatics alone, more powerful computers will have to be devised to organize the immense amount of information generated from these endeavors.

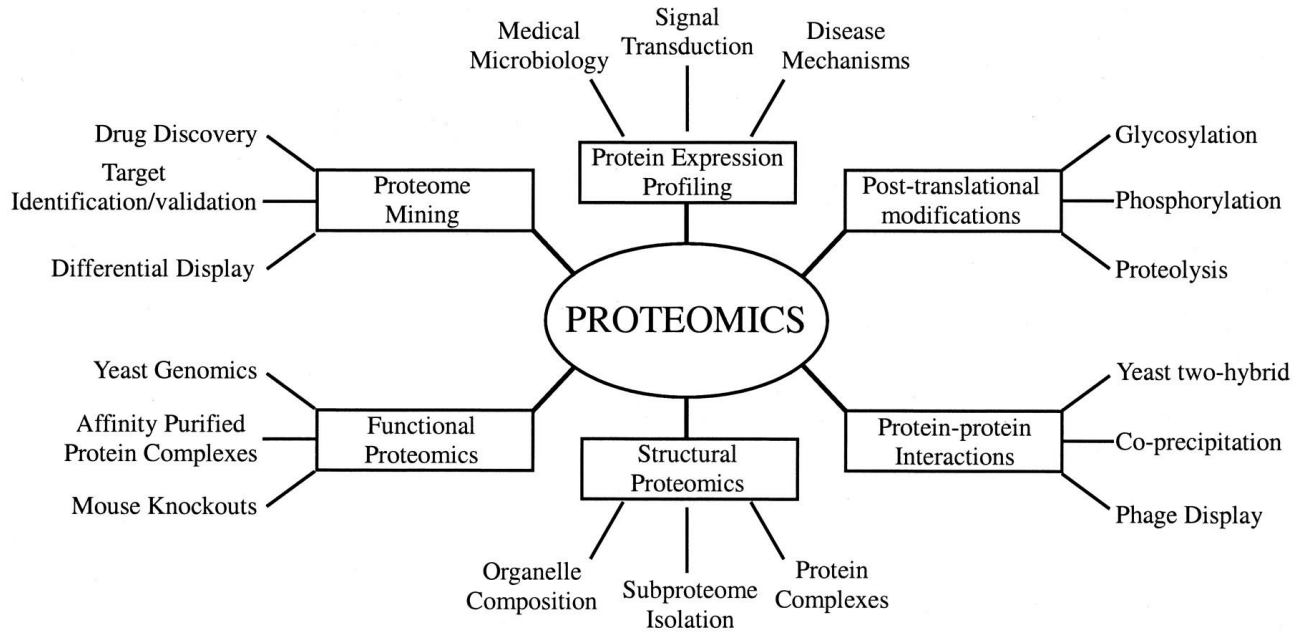


Figure. Types of proteomics and their applications to biology

- While *proteomics* generally refers to the large-scale experimental analysis of proteins, it is often specifically used for protein purification and mass spectrometry. Proteomics is increasingly taking up the position in the biological and biomedical research.

4.4.2 Transcriptomics and Metabolomics

1. Transcriptomics:

- Transcriptomics is the study of the transcriptome—the complete set of RNA transcripts that are produced by the genome, under specific circumstances or in a specific cell—using high-throughput methods, such as microarray analysis. Comparison of transcriptomes allows the identification of genes that are differentially expressed in distinct cell populations, or in response to different treatments.
- The initial product of genome expression is the transcriptome, a collection of RNA molecules derived from those protein-coding genes whose biological information is required by the cell at a particular time (*Figure*). These RNA molecules direct synthesis of the final product of genome expression, the proteome, the cell's repertoire of proteins, which specifies the nature of the biochemical reactions that the cell is able to carry out.

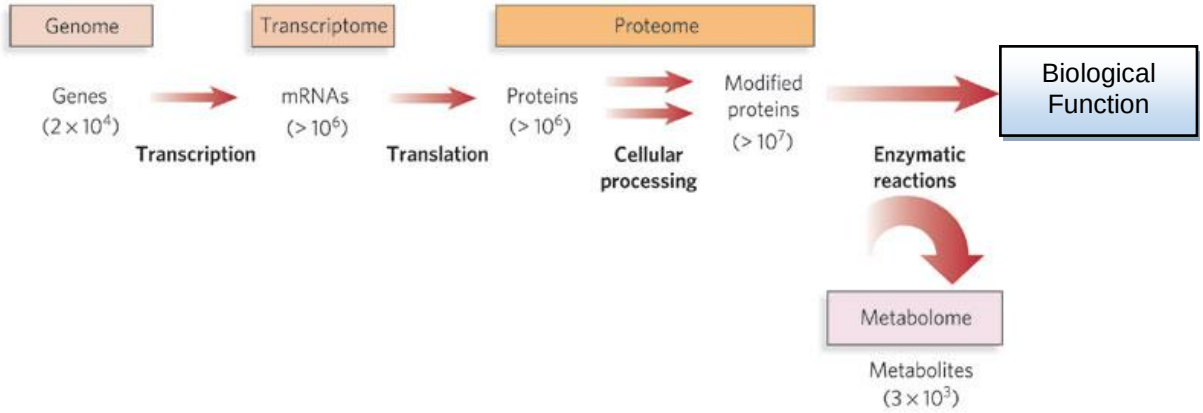
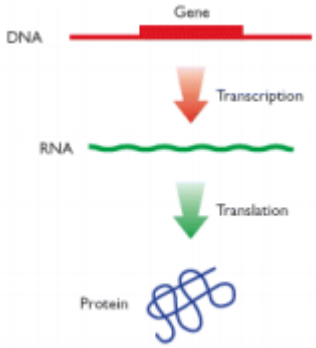


Figure. The relationship between genome, transcriptome and protome.

Transcriptome construction:

The transcriptome is constructed by the process called transcription, in which individual genes are copied into RNA molecules. Construction of the proteome involves translation of these RNA molecules into protein. Transcription and translation are important terms but it is unfortunate that the expression of individual genes is sometimes described simply as the two-step process 'DNA makes RNA makes protein'.



Applications:

- The transcriptomes of stem cells and cancer cells are of particular interest to researchers who seek to understand the processes of cellular differentiation and carcinogenesis.
- Analysis of the transcriptomes of human oocytes and embryos is used to understand the molecular mechanisms and signaling pathways controlling early embryonic development, and could theoretically be a powerful tool in making proper embryo selection in in vitro fertilisation.
- Transcriptomics is an emerging and continually growing field in biomarker discovery for use in assessing the safety of drugs or chemical risk assessment

2. Metabolomics:

- **Metabolomics** is the scientific study of chemical processes involving metabolites. Specifically, metabolomics is the "systematic study of the unique chemical fingerprints that specific cellular processes leave behind", the study of their small-molecule metabolite profiles.
- The metabolome represents the collection of all metabolites in a biological cell, tissue, organ or organism, which are the end products of cellular processes. mRNA gene expression data and proteomic analyses reveal the set of gene products being produced in the cell, data that represents one aspect of cellular function. Conversely, metabolic profiling can give an instantaneous snapshot of the physiology of that cell. One of the challenges of systems biology and functional genomics is to integrate proteomic, transcriptomic, and metabolomic information to provide a better understanding of cellular biology.

Metabolome:

- Metabolome refers to the complete set of small-molecule metabolites (such as metabolic intermediates, hormones and other signaling molecules, and secondary metabolites) to be found within a biological sample, such as a single organism.
- The word was coined in analogy with transcriptomics and proteomics; like the transcriptome and the proteome, the metabolome is dynamic, changing from second to second. Although the metabolome can be defined readily enough, it is not currently possible to analyse the entire range of metabolites by a single analytical method.
- For example, over 50,000 metabolites have been characterized from the plant kingdom, and many thousands of metabolites have been identified and/or characterized from single plants.

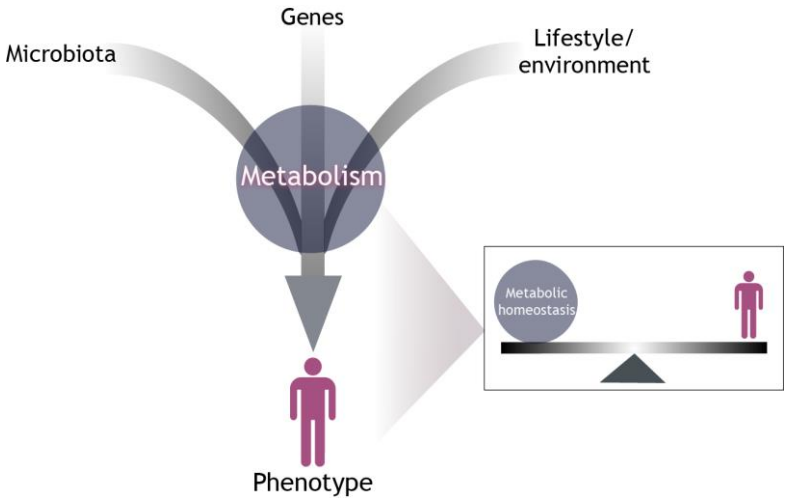


Figure. Metabolomics in science

- Although we did not devote attention to every area of biology or therapeutic area, the intent of this broad series was not only to convey how metabolomics can be used in a specific area of research (e.g. cancer), but actually, how metabolomics is a central science for interrogating any biological question (above figure).

Metabolomics Techniques:

a) Separation methods:

- Initially, analytes in a metabolomic sample comprise a highly complex mixture. This complex mixture can be simplified prior to detection by separating some analytes from others. Separation achieves various goals: analytes which cannot be resolved by the detector may be separated in this step; in MS analysis ion suppression is reduced; the retention time of the analyte serves as information regarding its identity. This separation step is not mandatory and is often omitted in NMR and "shotgun" based approaches such as shotgun lipidomics.
- Gas chromatography, especially when interfaced with mass spectrometry (GC-MS), is one of the most widely used methods for metabolomic analysis.
- High performance liquid chromatography (HPLC) is another common method for metabolomic analysis.
- Capillary electrophoresis (CE) has a higher theoretical separation efficiency than HPLC, and is suitable for use with a wider range of metabolite classes than is GC. As for all electrophoretic techniques, it is most appropriate for charged analytes.

b) Detection methods:

- Mass spectrometry (MS) is used to identify and to quantify metabolites after optional separation by GC, HPLC (LC-MS), or CE. GC-MS was the first hyphenated technique to be developed.
- Secondary ion mass spectrometry (SIMS) was one of the first matrix-free desorption/ionization approaches used to analyze metabolites from biological samples.
- Nuclear magnetic resonance (NMR) spectroscopy is the only detection technique which does not rely on separation of the analytes, and the sample can thus be recovered for further analyses. All kinds of small molecule metabolites can be measured simultaneously - in this sense, NMR is close to being a universal detector.

4.4.3 Proteomics techniques (2D PAGE)

Introduction:

- Two major approaches used are gel-based proteomics and “shotgun” proteomics. In the gel-based approach, proteins are resolved by electrophoresis or another separation method and protein features of interest are selected for analysis. This approach is best represented by the use of two-dimensional sodium dodecylsulfate polyacrylamide gel electrophoresis (2D-SDS-PAGE) to separate protein mixtures, followed by selection of spots.

2D-PAGE:

- 2D-PAGE is a form of gel electrophoresis in which separation and identification of proteins in a sample are done by displacement in 2 dimensions oriented at right angles to one another (orthogonal). This technique is also used to compare two or more samples to find differences in their protein expressions.

Steps in 2D-PAGE:

Basis for separation

- In this technique proteins are separated by two different physicochemical properties. In the first dimension proteins or polypeptides are separated on the basis of their net charges by isoelectric focusing and in the second dimension they are separated on the basis of their molecular masses by electrophoresis. Because it is unlikely that two molecules will be similar in both properties, molecules are more effectively separated in 2-D electrophoresis than in 1-D electrophoresis.

Sample preparation

- The goal of sample preparation is to solubilize maximum number of proteins and maintain their solubility throughout the process. The materials for sample should be carefully collected, snap frozen and ground under liquid nitrogen in the presence of protease inhibitors. After extracting proteins from source material they are then solubilized and denatured by means of chaotropes, detergents, and reducing agents. Hydrogen-bonds in the sample proteins are disrupted by chaotropes urea and thiourea. Uncharged detergents are used to disrupt hydrophobic interactions. Detergents such as CHAPS, Triton X-100, sulfobetaine SB3-10, and amidosulfobetaine are IEF-compatible additives. Disulfide bonds are reduced to sulfhydryls by reducing agents dithiothrietol (DTT), dithioerythritol, and tributyl phosphine (TBP).
- Sequential extraction is done to categorize proteins based on their solubility. This is an example of pre-fractionation to enrich low abundance proteins. Proteins are sequentially extracted into chaotrope/detergent solutions of increasing solubilization power. First, proteins are treated with an aqueous buffer, the insoluble proteins remaining from this extraction are treated with urea/CHAPS/TBP, and the insoluble proteins remaining from this step are then treated with urea/thiourea/CHAPS/SB 3-10/TBP. 2-D PAGE is then done to separate the proteins in each of the supernatants. The remaining insoluble material from the final extraction can be taken up in SDS-PAGE sample solution and run in a one-dimensional gel.

Isoelectric focusing (IEF)

- In IEF, proteins are separated by electrophoresis in a pH gradient based on their isoelectric point(pl). A pH gradient is generated in the gel and an electric potential is applied across the gel. At all pHs other than their isoelectric point, proteins will be charged. If they are positively charged, they will move towards the more negative end of the gel and if they are negatively charged they will move towards the more positive end of the gel. At its isoelectric point, since the protein molecule carry no net charge it accumulates or focuses into a sharp band.

Immobilized pH Gradient (IPG) and IEF run

- Immobilized pH gradients are used for IEF because the fixed pH gradients remain stable over extended run times at very high voltages. The pH gradients of IPGs are generated by means of buffering compounds that are covalently bound into polyacrylamide gels. IPGs are cast strips with plastic backing sheets and are commercially available in different pH ranges and lengths. They offer high resolution, great reproducibility, and allow high protein loads. Isoelectric focusing is run in the same solutions that are used to extract or solubalize the proteins. The IPG strips with the protein sample must be rehydrated in the rehydration/sample buffer during with protein samples are loaded into the strips. Rehydration can be active or passive. To load larger proteins active rehydration in small voltage is applied. After the run in IEF

cell, the proteins focus as bands on the strip according to their isoelectric points. The focused strips can be frozen for storage

SDS-PAGE;

IPG gel strips equilibration

- The proteins in the focussed IPG strips are uncharged because they are at their pI and so they will not move into the SDS- PAGE gel. So the strips are treated with SDS (sodium dodecyl sulfate), an anionic detergent which denatures the protein by breaking the disulfide bonds and gives negative charge to each protein in proportion to its mass. Without SDS, different proteins with similar molecular weights would migrate differently due to differences in folding, as differences in folding patterns would cause some proteins to better fit through the gel matrix than others. SDS linearizes the proteins so that they may be separated strictly by molecular weight. The SDS binds to the protein in a ratio of approximately 1.4 g SDS per 1.0 g protein (although binding ratios can vary from 1.1-2.2 g SDS/g protein), giving an approximately uniform mass:charge ratio for most proteins, so that the distance of migration through the gel can be assumed to be directly related to only the size of the protein. Proteins may be further treated with reducing agent, such as dithiothreitol (DTT) or TRP(Tributyl phosphine to break any reformed disulfide bonds and then alkalated with iodoacetamide to prevent reformation of disulfide bonds. A tracking dye like bromophenol blue may be added to the protein solution to track the progress of the protein solution through the gel during the electrophoretic run.

SDS-PAGE run

- The equlibrated IPG strip is placed on the top of the SDS-PAGE gel submerged in a suitable buffer and sealed in place with agarose gel. An electric current is applied across the gel, causing the negatively-charged proteins move out of the gel and migrate across the gel. Depending on their size, each protein move differently through the gel matrix. Smaller proteins travel farther down the gel, while larger ones remain closer to the point of origin. The proteins separate roughly according to size and therefore by molecular weight. It is common to run "marker proteins" of known molecular weight in a separate lane in the gel, in order to calibrate the gel and determine the weight of unknown proteins by comparing the distance traveled relative to the marker.

Visualization

- After electrophoresis the gel is stained to visualize the separated proteins. Commonly used stains are Coomassie Brilliant Blue or SYPRO Ruby or silver stain. different proteins will appear as distinct spot within the gel. Coomassie Brilliant Blue or SYPRO Ruby are compatible with Mass Spectrometry. Coomassie Brilliant Blue has detection limit about 10ng of proteins per spot and the gel images of spots can be captured by scanning densitometer which operate in visible light[[1]]. SYPRO Ruby can detect 1ng of proteins per spot and since it is fluorescent , the sopts are visualized by a fluorescent imager. Silver stain can detect spots containing proteins less than 1 ng and is the most sensitive non – radioactive protein visualization method.Laser devices for image capturing are useful for fluorescently stained gels.

Analysis

- The images can be further analyzed using image –analysis softwares. These softwares quantify proteins spots, match images and compare corresponding spots intensites of related gels, prepare gel data reports, remove background patterns, and integrate image information to databases. Alternately the proteins separated

can be obtained from the gel can be analyzed by MS for protein identification. 2D-PAGE can be used to study diffrential protein expression by comparing images from 2D –PAGE gels samples labeled with stable isotopes or fluorescent dyes.

Current uses of 2D electrophoresis in proteomics;

- Despite the now well-known limitations of 2D gels, which have been outlined above for membrane proteins and will be dealt with in more detail in 5 and 2D gels are still widely used in proteomics, and this roots in several key features.
- One of these features deals with the economy of proteomics. As shown on Figure, in 2D gel-based proteomics, the 2D gel part represents the essential workload of the whole process. It is at this step that the quantitative analysis is performed, and this quantitative analysis is usually used to perform spot selection.

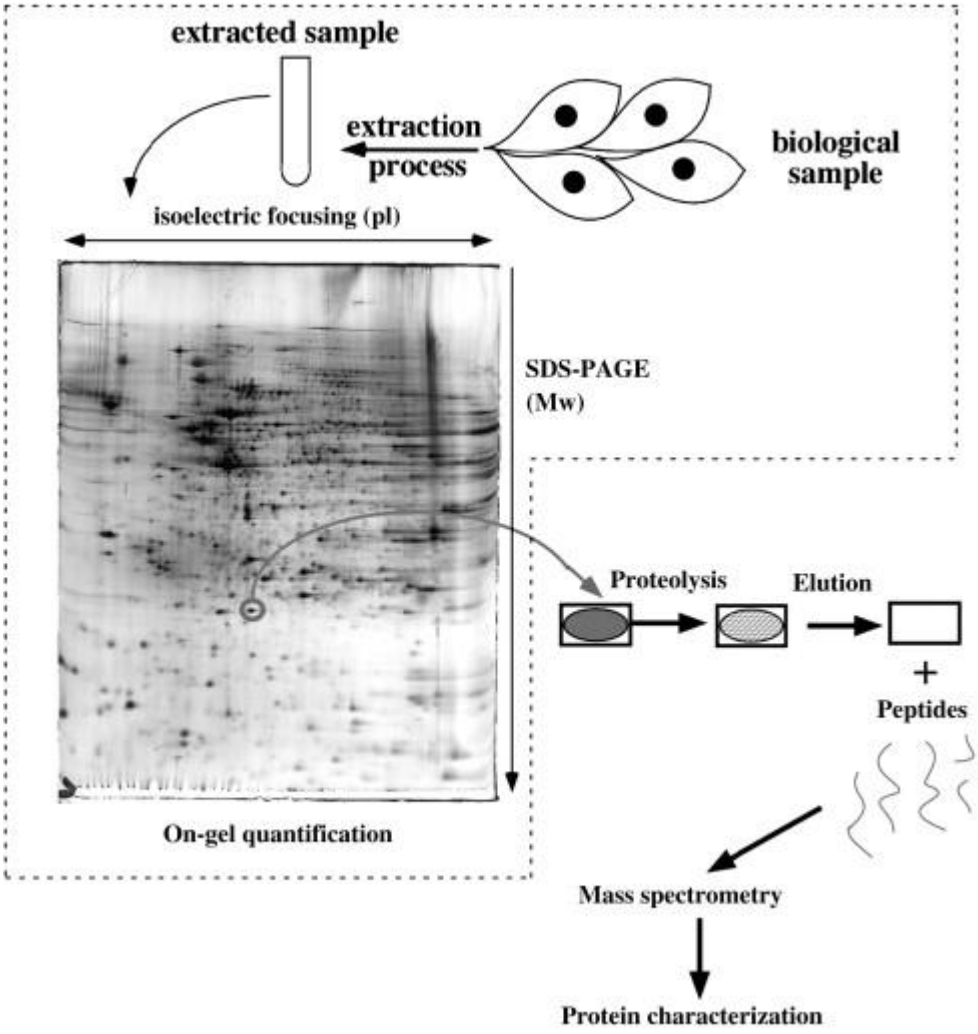


Figure: Scheme of principle of 2D gel-based proteomics.

- This has important consequences for the downstream mass spectrometry analysis. First of all, this means that only a very limited portion of the proteins present in the samples will need to be analyzed. This is especially true for a comparative study with replicates. If we imagine ten samples to be analyzed and compared, at 20 hours of mass spectrometry per sample, this represents in shotgun-type techniques 200 hours of MS. If the same analysis is carried out with 2D gels, at the end of the image analysis, maybe 20 different spots will be selected, and this represents at the very most 20 hours of mass spectrometry. This does not mean that 2D gel-based proteomics is more productive per se. It means that the burden put into the more

expensive, MS part is reduced in this scheme. However, when analyzing the comparative productivity of 2D gel-based proteomics and shotgun-type proteomics, it appears that the compared productivity of 2D gels over shotgun improves when the size of the sample series increases, due to the highly parallel nature of 2D gel-based proteomics.

4.5 Computational Biology – Multiple Sequence Analysis and Phylogenetic alignment.

4.5.1 Introduction:

- *Computational biology* involves the development and application of data-analytical and theoretical methods, mathematical modelling and computational simulation techniques to the study of biological, behavioural, and social systems.[1] The field is broadly defined and includes foundations in computer science, applied mathematics, animation, statistics, biochemistry, chemistry, biophysics, molecular biology, genetics, genomics, ecology, evolution, anatomy, neuroscience, and visualization.
- Computational biology is different from biological computation, which is a subfield of computer science and computer engineering using bioengineering and biology to build computers, but is similar to bioinformatics, which is an interdisciplinary science using computers to store and process biological data.

4.5.2 Multiple Sequence Analysis

- A **multiple sequence alignment** (**MSA**) is a sequence alignment of three or more biological sequences, generally protein, DNA, or RNA. In many cases, the input set of query sequences are assumed to have an evolutionary relationship by which they share a lineage and are descended from a common ancestor.

- In this example multiple sequence alignment is applied to a set of sequences that are assumed to be homologous (i. e., having a common ancestor). Homologous nucleotides are placed in the same column of a multiple alignment. Note the phylogenetic tree attached to the left.
- While multiple sequence alignment is natural generalization of pairwise sequence alignment, there are lots of new questions and issues concerning the scoring function, the significance of scores, the gap penalties, and efficient implementations.

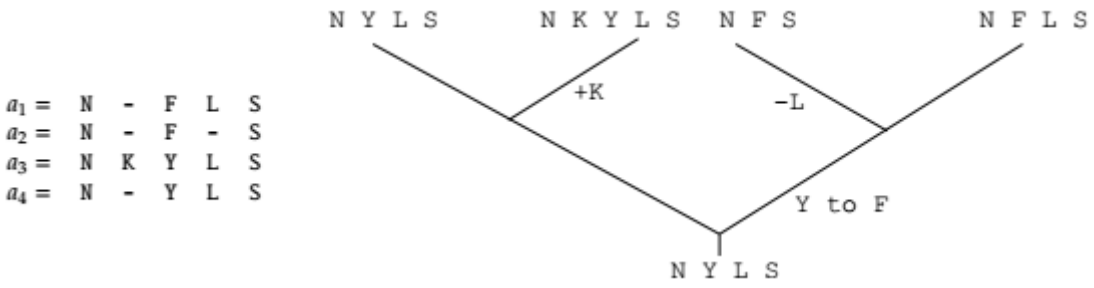
- From the resulting MSA, sequence homology can be inferred and phylogenetic analysis can be conducted to assess the sequences' shared evolutionary origins. Visual depictions of the alignment as in the image at right illustrate mutation events such as point mutations (single amino acid or nucleotide changes) that appear as differing characters in a single alignment column, and insertion or deletion mutations (indels or gaps) that appear as hyphens in one or more of the sequences in the alignment. Multiple sequence alignment is often used to assess sequence conservation of protein domains, tertiary and secondary structures, and even individual amino acids or nucleotides.

A standard multiple alignment program: ClustalW

- ClustalW (Thompson, Higgins & Gibson, 1994) is one of the standard programs implementing one variant of the progressive method in wide use today for multiple sequence alignment. The W denotes a specific version that has been developed from the original Clustal program.
- The basic steps of the algorithm implemented in ClustalW are:
  1. Compute the **pairwise alignments** for all against all sequences. The similarities are stored in a matrix (sequences versus sequences).
  2. Convert the sequence similarity matrix values to **distance** measures, reflecting evolutionary distance between each pair of sequences.
  3. Construct a tree (the so-called **guide tree**) for the order in which pairs of sequences are to be aligned and combined with previous alignments. This is done using a neighbour-joining clustering algorithm. In the case of ClustalW, a method by Saitou & Nei is used.
  4. **Progressively align** the sequences/alignments together into each branch point of the guide tree, starting with the least distant pairs of sequences. At each branch point, one must do either a sequence-sequence, sequence-profile, or profile-profile alignment.

MSA use in phylogenetics:

- Multiple sequence alignments can be used to create a phylogenetic tree. This is made possible by two reasons. The first is because functional domains that are known in annotated sequences can be used for alignment in non-annotated sequences. The other is that conserved regions known to be functionally important can be found. This makes it possible for multiple sequence alignments to be used to analyze and find evolutionary relationships through homology between sequences. Point mutations and insertion or deletion events (called indels) can be detected.
- Multiple sequence alignments can also be used to identify functionally important sites, such as binding sites, active sites, or sites corresponding to other key functions, by locating conserved domains. When looking at multiple sequence alignments, it is useful to consider different aspects of the sequences when comparing sequences. These aspects include identity, similarity, and homology. Identity means that the sequences have identical residues at their respective positions. On the other hand, similarity has to do with the sequences being compared having similar residues quantitatively. For example, in terms of nucleotide sequences, pyrimidines are considered similar to each other, as are purines. Similarity ultimately leads to homology, in that the more similar sequences are, the closer they are to being homologous. This similarity in sequences can then go on to help find common ancestry.
- One main application of multiple sequence alignment lies in phylogenetic analysis. Given an MSA, we would like to reconstruct the evolutionary tree that gave rise to these sequences, e.g.:



We can find out how the positions of the sequences correspond to each other.

4.5.3 Phylogenetic Alignment

A) <u>Definitions</u>:

- **Phylogenetics:**
- In biology it is the study of the evolutionary history and relationships among individuals or groups of organisms (e.g. species, or populations). These relationships are discovered through phylogenetic inference methods that evaluate observed heritable traits, such as DNA sequences or morphology under a model of evolution of these traits. The result of these analyses is a phylogeny (also known as a phylogenetic tree) – a hypothesis about the history of evolutionary relationships.

- **Phylogenetic tree** :
- Phylogenetic tree also known as “evolutionary tree” is the graphical representation of the evolutionary relationship between the taxa/genes in question.

- A **dendrogram** is a broad term for the diagrammatic representation of a phylogenetic tree.

- The **cladogram** is a dendogram which explains only genealogy of the taxa but says nothing about the branch lengths or time periods of divergence.

- The **phylogram** (additive tree) is a phylogenetic tree that explicitly represents a number of character changes (nucleotide/amino acid changes/number of character variations) through its branch lengths. In case of phylogram the evolutionary distance between any two taxa is given by sum of the branch lengths connected them. Though these trees may be rooted or unrooted, often these trees lack a root.

B) <u>Process of construction of the phylogenetic tree</u>

- The whole process of construction of the phylogenetic tree is divided into five different steps, viz.

  Step 1: Choosing an appropriate markers for the phylogenetic analysis
  Step 2: Multiple sequence alignments
  Step 3: Selection of an evolutionary model
  Step 4: Phylogenetic reconstruction
  Step 5: Evaluation of the phylogenetic tree

  Step 1: Choosing an appropriate markers for the phylogenetic analysis
  - Any biological information that can be used to infer the evolutionary relationship among the taxa is known as a phylogenetic information marker.

- It can be anything like DNA, RNA, protein, RFLP, AFLP, ISSR, allozymes, and conserved intronic positions, etc.
- Identification of conserved genetic loci (coding- or noncoding) is the first step in analyzing the phylogenetic relationship.
- Both coding (genes) and non-coding genetic region can be used for the analysis of phylogenetic relationships.
- However, selected sequence(s) must satisfy the defined necessary rules:
  a. The sequence should have a long evolutionary history of conservation, as this feature facilitates, firstly in the preservation of long evolution-selection episodes, and secondly, aids in easy amplification of the target sequences from distant taxa.
  b. Conserved, slow evolving genes may be used to resolve the evolutionary relationship between distantly related species while fast evolving genes should be choose for the recently evolved species or intra-species.
  c. Amino acid sequences are more informative while inferring the evolutionary relationship among distantly related taxa, and conversely, nucleotide information for recently evolved/closely related species.
  d. The sequences need to be employed in the phylogenetic analysis should be tested for their usability in a given lineage (for instance, mitochondrial (cytochrome C oxidase subunit I & II (CoxI & II)), chloroplast (trnH-psbA, matK, rpoC, rpoB, rbcL), and nuclear (16S ribosomal RNA) conserved genes are preferred to use for analyzing animal, plant, and microbial species, respectively- and are called “barcode genes”).
  e. Finally, if, objective is to estimate the divergence periods between taxa, the selected gene or protein sequences should essentially follow the molecular clock hypothesis. However, recently relaxed molecular clock models have also been proposed. This step follows successful polymerase chain reaction amplification of the target gene/protein, followed by sequencing and editing of the sequences for further analysis.

Step 2: Multiple sequence alignments

- The main aim of multiple sequence alignment is to compare the three or more nucleotide or protein sequences and to provide the basis for calculation of the sequence diversities/divergences to infer the evolutionary relationship among the taxa.
- Different models (discussed below) have been proposed based on various assumptions to calculate the sequence divergences between the sequences or taxa. Hence, the correct sequence alignment is mandatory in order to get the true phylogeny that is representative of the evolutionary relationship among the taxa.

Step 3: Selection of an evolutionary model

- Evolutionary models are sets of assumptions about the process of nucleotide or amino-acid substitution. They describe the different probabilities of change from one nucleotide or amino acid to another, with the aim of correcting for unseen changes along the Phylogeny.

Step 4: Phylogenetic reconstruction

- Two different methodologies are employed by the presently available programs to generate the dendograms;
  a. clustering methods-where two most closely related taxa are placed under single inter-node and further add third taxa considering within internodes taxa as a

single group. In this way, the program progressively adds the other remaining taxa to yield final phylogenetic tree

b. second type of methods generate the 'n' number of trees proportional to the number of taxa involved in the phylogenetic analysis followed by the selection of best fit tree topology (increased likelihood or probability) for a given evolutionary model.

Step 5: Evaluating the phylogenetic tree

- This process can be performed using two evaluation methods, namely bootstrap method and interior-branch test.

  - In phylogenetics, ***bootstrapping*** is conducted using the columns of the character matrix. Each pseudoreplicate contains the same number of species (rows) and characters (columns) randomly sampled from the original matrix, with replacement. A phylogeny is reconstructed from each pseudoreplicate, with the same methods used to reconstruct the phylogeny from the original data. For each node on the phylogeny, the nodal support is the percentage of pseudoreplicates containing that node.

  - ***Inferior-branch test*** is a t-test, which is computed using the bootstrap procedure, is constructed based on the interior branch length and its standard error and is available only for the neighbor-joining and Minimum Evolution trees. MEGA (molecular evolutionary genetics analysis) shows the confidence probability in the Tree Explorer; if this value is greater than 95% for a given branch, then the inferred length for that branch is considered significantly positive. Select test of phylogeny for either of these trees in the Analysis Preferences dialog.

***